Conservation and synteny of long non-coding RNAs invertebrate genomes and their identification in novel transcriptomes

Thesis

Conservation and Synteny of Long Non-Coding RNAs in Vertebrate Genomes and their Identification in Novel Transcriptomes

A thesis submitted to the Open University of London for the degree of

Doctor of Philosophy

by

Swaraj Basu

Master of Science in Biotechnology and Bioinformatics, SRM University, Chennai – Tamil Nadu, India

Stazione Zoologica Anton Dohrn, Naples, Italy

The Open University, London, United Kingdom

Director of studies: Dr. Remo Sanges, Ph.D.

External Supervisor: Dr. Ferenc Müller, Ph.D.

Co-supervisor: Dr. Euan Robert Brown, Ph.D.

December, 2013

Abstract

Long non-coding RNAs (lncRNAs) are a biological entity defined by *what they are not*, rather than by *what they are*. This indicates that our knowledge about them is sensibly limited. The aim of my PhD is to gain insights into the evolution and the functions of lncRNAs through computational approaches and the usage of large scale functional genomics dataset. I developed an annotation pipeline, which can effectively identify lncRNAs in entire transcriptomes. The pipeline is able to accurately annotate the coding genes while predicting a conservative estimate of the lncRNA population. It allowed me to show, for the first time, the presence of lncRNA transcription in a diverse range of organisms. Further, I analysed sequence and positional conservation of lncRNAs, demonstrating the presence of short segments of conserved sequence in lncRNAs and the existence of several syntenically conserved non-coding transcripts over large evolutionary distances. However, I also demonstrate that positional conservation of lncRNAs with a flanking coding gene is generally independent from the conservation of the lncRNA expression with respect to the coding gene. Finally, I have characterised the diversity of lncRNA transcription in specific cells and developmental stages of two teleost fishes. In summary, the work presented in the thesis provides novel findings and contributions in the field of lncRNAomics.

Dedication

I dedicate this thesis to my grandfather, Sukumar Bose (1914 -1987). I have little but cherished memories of my time spent with him.

Acknowledgements

It is going to be a long acknowledgement. I would like to thank the Open University (OU) and Stazione Zoologica Anton Dohrn for giving me the opportunity and fellowship to pursue my Ph.D. I also want to thank Dr. Raffaella Casotti (coordinator OU Ph.D program) for her support throughout the duration of my Ph.D. I must specially thank Gabriella Grossi (secretary OU Ph.D program), for providing excellent help in resolving various Ph.D program related queries and timely reminders of important academic deadlines. I express my sincere gratitude to my thesis defense committee members, Dr. Derek Stemple, Dr. Paolo Sordino, and Dr. Marina Montresor for letting my defense be an enjoyable moment and for their brilliant comments, questions and suggestions.

I would like to express a special appreciation for my external supervisor Dr. Ferenc Müller whose excellent supervision, criticism and biological inference of my computational predictions aided in my growth as a researcher and shaped the thesis in its final form. The three months spent in his laboratory was a beneficial and learning experience for me. I spent a productive period of my Ph.D tenure working with Yavor, Irene, Jennifer, Harmeet, Emma, Padma and Nicholas in Birmingham. Specially I would like to thank Yavor, Emma and Irene for being patient while explaining me the basics of molecular biology. I sincerely appreciate Irene's hard work in generating the zebrafish islet cell data and her efforts towards the experimental validation of lncRNAs.

I would like to thank my co-supervisor Dr. Euan Robert Brown for the encouragement given to me during my PhD tenure. It is difficult to put into words the immense contribution of Dr. Remo Sanges, my director of studies, in my ability to think critically, design experiments, write scientifically and finally complete the thesis. I would just like to say "thank you Remo for being patient with me during all the times when I was lazy, delving in my science fiction land or simply lacking enthusiasm and motivating me at the right moments".

I really appreciate the calm and patience shown by my parents during all these years when I have stayed away from home. Their support has always made me confident that I am on the right track in my life. I sincerely appreciate the support received from my relatives (Kabita Ghosh, Tarun Ghosh, Leena Francis, Harry Francis, Shuchita Lokho), my sister and my cousins (Dia, Pooja, Justin and Sania) during my Ph.D tenure. My friends from childhood and college have played a big role in helping me remain composed and focused during my Ph.D period. Sandeep (Chanda), Ramesh, Satyam, Sushmita (Sharma), Shuchita, Surya, Swarna, Abhimanyu, Ranjeeta, Jeetendra (Gujrat) and Nitesh you guys will always be special to me, thanks for all the support and encouragement.

I would like to thank Dr. Subhra Chakraborty for giving me the opportunity to work in NIPGR. I have spent the best days of my stay at NIPGR, Delhi working with Nasheeman di. Thanks for your constant support and encouragement. My dear friends at NIPGR, I owe all of you for making the time spent there so

memorable specially Rajul, Vinay and Sudip. Each one of you have an important space in my life, which cannot be replaced by anyone else. Irfan, Manindra, Anand, Upendra, Mukesh, Manmohan, Rajiv and Jeetendra I have fond memories of all of you, thanks for being such great company.

A big part of the duration of my Ph.D has been a wonderful experience because of my friends and the charismatic city of Napoli. Vasco, Gianluca Santamaria, Davide, Heather, Nicole, Cecilia, Katharina, Alessandra, Evegeniya, Claudia (Racioppi and Cuomo), Laura (Vitale), Elisabetta, Shrikant, Raghu and Sneha thanks for all your kind gestures, affection and love. The dinners organised with Deepak and Gauri remain an important part of the Napoli experience and the happy memories of those days will always remain close to me. I owe a special gratitude to my dear lab members Giuseppe and Veerendra, our little conversations, coffee breaks and scientific discussions always replaced stress with motivation and joy.

I would like to add a few lines about a few friends of mine who have influenced my life in a positive way over the years. I must really thank Nisha for helping me grow as an individual and making me overcome my insecurities both by her presence and later absence. Shubhendu, Dipto, Manish and Manu have been my pillars of strength during the toughest period of my life. Your encouragement and belief in me is the reason I never lost the faith in myself. Suncica and Krzysztof, both of you have remained the patient ear with whom I can reliably share my joys

and woes. Sincere thanks for never ever deciding to throw a brick on my head after listening to my stories. Eleonora and Laura (Escalera), both of you are special-special to me, thanks for being such great friends always there for support. Ashwani and Francesco, thanks for being such awesome friends and brothers. I have always enjoyed our conversations and arguments, they have given me perspective and made me less critical of life. Apart from my supervisors, friends and colleagues a special note of thanks goes to the authors of grep, sort and uniq commands which I have used infinite times during my Ph.D. "Things change. And friends leave. Life doesn't stop for anybody – *Stephen Chbosky, The perks of being a wallflower*". I must add that friends who care don't leave how annoying the situation might get, and the presence of such friends have encouraged me to keep going at all times.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABCC8	ATP-binding cassette, sub-family C (CFTR/MRP), member 8
AIRN	Antisense of IGF2R non-protein coding RNA
ALDH1	Adehyde dehydrogenase 1
ANCR/DANCR	Angelman syndrome chromosome region
AUC	Area Under Curve
BACE1	β-secretase enzyme, beta-site APP cleaving enzyme-1
BACE1-AS	BACE1 antisense transcript
BLAST	Basic Local Alignment Search Tool
BMP4	Bone Morphogenetic Protein 4
CAGE	Cap Analysis of Gene Expression
CCG	Closest Coding Gene
cCNS	Conserved mouse Central Nervous System lncRNA
cNCNS	Conserved mouse non Central Nervous System lncRNA
Censembl	Conserved mouse non Ensembl lncRNA
CDD	Conserved Domain Database
CDK5	Cyclin-dependent kinase 5
Cdkn1a	Cyclin dependent kinase inhibitor
CNEs	Conserved Non-coding Elements
CONC	Coding Or Non-Coding
CPAT	Coding Potential Assessment Tool
CPC	Coding Potential Calculator
CRNDE	Colorectal Neoplasia Differentially Expressed
CTCF	CCCTC-Binding Factor
CTNNB1	Catenin (cadherin-associated protein) Beta 1
DAVID	Database for Annotation, Visualization and Integrated Discovery
DKK-1	Dickkopf WNT signaling pathway inhibitor 1
Dnmt3a	DNA methyltransferase
Dvl3	Dishevelled segment polarity protein 3
eIF4A	eukaryotic translation Initiation Factor 4A1
ENCODE	ENCyclopedia Of DNA Elements
ESTs	expressed sequence tags
ETF1a	Eukaryotic Translation Termination Factor 1a
EZH2	Enhancer Of Zeste Homolog 2

FANTOM	Functional Annotation of the Mouse
FDR	False Discovery Rate
FoxA2	Forkhead box protein A2
FPKM	Fragments per Kilobase of exon per Million
GAS5	Growth Arrest Specific 5
GCK	Glucokinase
GLIS3	GLIS family zinc finger
GO	Gene Ontology
GR	Glucocorticoid Receptor
GRBs	Genome Regulatory Blocks
GTF	Gene Transfer Format
Jpx	XIST activator
H3K27ac	Histone 3 lysine 27 acetylation
H3K4Me1	Histone 3 lysine 4 monomethylation
H3K4Me3	Histone 3 lysine 4 trimethylation
HAR1A	Human Accelerated Region 1A
HAR1B	Human Accelerated Region 1B
HAR1F	Human Accelerated Region 1F
HCEs	Highly Conserved Elements
HCNEs	Highly Conserved Non-coding Elements
Hdacs	Histone deacetylases
HEN2	Nescient Helix Loop Helix 2
HNF1a	Hepatic Nuclear Factor 1
HNF1A-AS1	Hepatic Nuclear Factor 1A antisense 1
hnRNAs	heterogeneous nuclear RNAs
HOTAIR	Hox transcript antisense RNA
HOTAIRM1	Hox Antisense Intergenic RNA Myeloid 1
HOTTIP	Hoxa distal transcript antisense RNA
HoxA	Homeobox A
HSP	Highest Scoring Pair
HULC	Highly Upregulated in Liver Cancer
HuR	Human Antigen R
IGF2	Insulin-like Growth Factor 2
iNOS	Nitric Oxide Synthase
INS	Insulin
Insl5B	Insulin like 5b

Isl1	ISL LIM homeobox 1
Isl2	ISL LIM homeobox 2
KEGG	Kyoto Encyclopedia of Genes and Genomes
Klf6	Kruppel-like factor 6
LCoR	Ligand Dependent Nuclear Receptor Corepressor
fabp1a	Fatty Acid Binding Protein 1, Liver
LincIGs	intergenic regions which contain a lincRNA
LincRNA-MD1	Long non coding RNA Muscle Differentiation 1
LincRNAs	Long intergenic non-coding RNAs
LMO3	LIM Domain Only 3
LncRNA	Long non-coding RNA
LRLSLDA	Laplacian Regularized Least Squares for LncRNA-Disease Association
LSD1	Lysine K-Specific Demethylase 1A
MAF	Multiple Alignment Format
MALAT1	Metastasis Associated Lung Adenocarcinoma Transcript 1
Maml1	Mastermind-like 1
MEF2C	Myocyte Enhancer Factor 2C
MIAT	Myocardial Infarction Associated Transcript
myl7	Myosin, light chain 7
MZT	Maternal to Zygotic Transition
NCP	Non-coding potential
ncRNA-a	ncRNA-activating
ncRNAs	Non-coding RNAs
Nctc1	Non coding transcript 1
NEAT1	Nuclear Paraspeckle Assembly Transcript 1
NF-Y	Nuclear transcription Factor Y
oCNEs	Olfactores conserved non-coding element
Oct4	Octamer-Binding 4
ORF	Open Reading Frames
PANDA	P21 Associated NcRNA DNA damage Activated
PAX4	Paired box gene 4
Pcdh2g16	Protocadherin 2 gamma 16
Pcdha9	Protocadherin Gamma Subfamily A, 9
Pcdhg	Gamma protocadherins
PDX1	Pancreatic and duodenal homeobox 1

PHLPP1	PH domain and Leucine rich repeat Protein Phosphatases
pIdentity	Percentage identity of aligment
PLoNCs	Potential Long Non-Coding Sequences
PRC2	Polycomb Repressive Complex 2
QAlength	query alignment length
QCoverage	query coverage
ROC	Receiver Operating Characteristic
RPKM	Reads per Kilobase per Million
Rux2	Roughex 2
SAMD11	Sterile Alpha Motif Domain containing 11
SF1	Splicing Factor 1
six3b	Six Homeobox 3
SLC2A2	Solute carrier family 2
Slit1b	Slit homolog 1
Smg1	Smaug 1
SNHG15	Small Nucleolar RNA Host Gene 15
snoRNA26	Small Nucleolar RNA 26
Sox2ot	Sox2 overlapping transcript
Sox9	SRY (sex determining region Y)-box 9
SRA	Steroid receptor RNA activator 9
SVM	Support Vector Machine
SYT8	Synaptotagmin VII
T2DGADB	Type 2 Diabetes Genetic Association Database
T2DM	Type 2 diabetes mellitus
TEs	Transposable Elements
Tsix	Xist Antisense RNA
UCEs	Ultra Conserved Elements
Uchl1	Ubiquitin carboxyl-terminal esterase L1
UCRs	Ultra Conserved non-coding genomic Regions
Usp46	Ubiquitin specific peptidase 46
vlincRNAs	very long intergenic RNAs
VMLRs	Vertebrate microsyntenic lincRNAs with retained orientation
VMLs	Vertebrate microsyntenic lincRNAs
wlincIGs	Intergenic regions without lincRNAs
Wnt	Wingless-type MMTV integration site
XCI	X Chromosome Inactivation

Xi	Inactivated X chromosome
Xic	X inactivation center
Xist	X inactive specific transcript
YTHDF1	YT521-B homology Domain Family 1
zCNS	Conserved zebrafish segment of mouse CNS lncRNA
zNCNS	Conserved zebrafish segment of mouse NCNS lncRNA
zEnsembl	Conserved zebrafish segment of mouse Ensembl lncRNA
zPHS	PhastCons elements for zebrafish

Chapter 1

General Introduction

1.1 The history of the "dark matter"

1.1.1 Pervasive transcription in the eukaryotic genome

The idea of non-coding transcription in the eukaryotic genome took seed a few decades ago with reports of 50% heterogeneous nuclear RNAs (hnRNAs) containing non-coding sequences (Holmes et al., 1972; Pierpont and Yunis, 1977). The hnRNAs are transcribed from heterochromatic, repetitive and non-repetitive regions in the mammalian genomes. This gave an impression that the transcribed part of the genome is more than what is credited to protein-coding genes, ribosomal RNAs and transfer RNAs. The discovery of small nuclear and small nucleolar RNA (Reddy et al., 1979; Rein, 1971) and their role in RNA processing failed to interest the scientific community into looking further in the non-coding genome. The late 1990's and the early 2000's witnessed the advent of microarray and sequencing technologies which proved to be powerful tools in measuring the transcriptional output of mammalian genomes. A conspicuous observation made with the aid of these large-scale technologies was the widespread transcription in the mouse (Carninci et al., 2005; Okazaki et al., 2002) and human genomes (Bertone et al., 2004; Cawley et al., 2004; Ota et al., 2004; Rinn et al., 2003). The

phrase *"pervasive transcription of the genome"* gained prevalence from these reports and the term *"dark matter"* was coined for all transcribed genomic regions not localized on a protein coding region (Johnson et al., 2005). To fathom the diversity and complexity of the human genome "The ENCyclopedia Of DNA Elements (ENCODE) Project" was launched aiming to identify all functional elements in the human genome sequence (The ENCODE Project Consortium, 2004). The pilot phase of ENCODE focused on a 30 megabase region (1%) of the human genome, reporting such a pervasive transcription that the majority of bases in the analyzed regions were associated with at least one primary transcript (Birney et al., 2007). A question which inevitably arises on acknowledging this rife transcriptional activity is whether it has a biological significance.

1.1.2 Functionality of the *"dark matter"* against the *"transcriptional noise"* hypothesis

At this point of time it was known that much of the *"dark matter"* (non-coding) transcription occurred at very low levels, thus making it difficult to detect them with the available technologies (Kapranov et al., 2002). Hence opponents of the argument termed it as *"transcriptional noise"* due to insufficient progress in demonstrating the usability of non-coding transcripts (Hüttenhofer et al., 2005). Support for the *"transcriptional noise"* hypothesis came from independent reports which elucidated various aspects of the eukaryotic transcription. Firstly a report suggested that majority of the eukaryotic transcription initiation events by RNA polymerase II are not associated with a functional transcript and hence represent

"transcriptional noise" (Struhl, 2007). Another study demonstrated that a region of intense transcriptional activity (coding regions) in mammalian genomes may frequently lead to a transcriptional ripple effect marking non-specific transcription in the surrounding area (Ebisuya et al., 2008). Finally eukaryotic transcription factors were reported to have wide spread binding sites but dependent on clustering of binding sites for their specificity of action (Wunderlich and Mirny, 2009). Hence Wunderlich *et al* proposed, that the majority of binding events by eukaryotic transcription factors were non-functional. Further mammalian intergenic unannotated transcripts were reported to show a tendency to lie near coding genes and predicted to be alternative exons, promoter/terminator-associated RNAs or pre-mRNA fragments of coding genes (van Bakel et al., 2010), thus indicating an over-estimation of bona fide intergenic transcripts reported by previous studies. The results by van Bakel were contested in another study, which stated the lack of sequencing depth and poor transcript assembly as the reason for non-detection of lowly expressed novel intergenic transcripts (Clark et al., 2011). Clark *et al* further argued that the sequencing by van Bakel took into account only polyadenylated RNA while a large proportion of the novel intergenic transcripts may not be polyadenylated. The choice of a complex tissue like brain was cited as another reason for the inability of van Bakel *et al* to detect lowly expressed, highly tissue specific intergenic transcripts. The objections raised by Clark *et al* have support from a previously published study which reported the presence of novel intergenic non-coding transcripts expressed in specific cell types revealing the dynamic nature of the cells transcriptional machinery (Guttman et al., 2010). The

strife on the functionality of the *"dark matter"* paved the way for two prominent schools of thought. Opponents of *"pervasive transcription"* believed that the *"dark matter"* might comprise of an intricate transcript population comprising of a small fraction of the total cellular RNA content. Thus the *"dark matter"* is proposed to be an offshoot of cellular processes, justifying its label as *"transcriptional noise"* (van Bakel and Hughes, 2009; van Bakel et al., 2010; Struhl, 2007). A contradictory view proposes that a relevant fraction of the total cellular RNA might be comprised of the *"dark matter"*, which makes it an essential component for organism development and differentiation. Further the absence of suitable technology to measure the relative RNA content was cited to be the principle reason for the *"dark matter"* to be considered as *"transcriptional noise"* (Kapranov et al., 2007a; Mattick, 2011; St Laurent and Wahlestedt, 2007).

1.1.3 Estimation of the *dark matter* abundance with next-generation sequencing technologies

A plausible answer to the suppositions came with the onset of next-generation sequencing technologies, specifically RNA-seq to measure the transcriptional output of a cell population, tissue or whole organism (Cloonan and Grimmond, 2008; Mortazavi et al., 2008; Wang et al., 2009) at an extremely high depth. This technology is based on shearing of the total RNA content from a cell tissue or organism followed by its ligation to adapters, PCR amplification and sequencing. The sequencing gives a digital count of the number of reads (25-150 bases), each read representing a small fragment of a transcribed RNA from the initial

population. Thus the count of all reads belonging to a particular gene, genomic feature or the total *"dark matter"* gives its fraction in the total sampled RNA population. Although simplistic in approach initial reports gave a motley set of figures from as low as 7% (Mortazavi et al., 2008) to a maximum of 40-50% (Cloonan et al., 2008; Morin et al., 2008) of non-coding transcription. The study by van Bakel *et al* followed these reports to estimate that *"dark matter"* comprises 12% of total polyadenylated RNA in human and mouse cells (van Bakel et al., 2010). An important aspect of the above mentioned studies was the preparation of RNA for sequencing. The proclivity of current sequencing technologies to alter the initial RNA population during reverse transcription, adapter ligation, library amplification and PCR could be a major factor in obtaining diverse estimates of the "dark matter" (Aird et al., 2011; Mamanova et al., 2010; Shiroguchi et al., 2012). An alternative approach could be the use of a sequencing technology which does not rely upon cloning, amplification or ligation of RNA molecule such as single molecule sequencing which, in principle, uses a high fidelity DNA polymerase coupled with flourescence microscopy to obtain the sequences at single base pair resolution (Braslavsky et al., 2003; Pushkarev et al., 2009). Indeed, single RNA deep sequencing showed that ~50% of human transcriptome is *"dark matter"* (non-ribosomal, non-mitochondrial unannotated transcripts of unknown function) (Kapranov et al., 2010). This study highlighted the inability of sequencing technologies to detect diverse RNA classes due to a bias towards amplifying polyA RNA and in turn coding transcripts which comprise of a significant fraction of polyA RNA (Kapranov et al., 2010). Another study involving in depth tiling arrays

of targeted genomic regions reported the presence of a complex repertoire of novel intergenic non-coding transcripts with low expression levels (Mercer et al., 2011). This indicates that a low signal in the tiling array may represent a complex transcript population of low abundance. Mercer *et al* proposed that such transcripts could emanate from a very specialized cell type hence their low levels of expression cannot be held as a confirmation of them being transcriptional noise. At this point an important issue which remained unexplored was that the transcription of repetitive regions are refrained from experiments like tiling arrays due to non-specific binding (Kapranov et al., 2007b). Although it is a work in progress, technologies like Cap analysis of gene expression (CAGE) (Kodzius et al., 2006) allowed the accurate measurement of repeat transcription in multiple human cell lines (Faulkner et al., 2009). A significant fraction of transcribed elements were observed to fall in repetitive regions of the genome and there was enrichment of repeat expression in embryonic tissues. The era of arguments and counter-arguments in defense of pervasive transcription led to few conclusions drawn with unanimous acceptance:

- Much of the eurkaryotic transcription is concentrated around coding genes.
- Widespread transcription, specificity of expression and complexity in the population characterise the non-coding RNAs in the cell.
- Non-coding RNAs (ncRNAs) are lowly expressed as compared to coding genes.

1.1.4 The new face of *dark matter*: Long non coding RNAs

Articulating on the widespread transcription of ncRNAs directs us towards their population subtypes. Multiple classes of ncRNAs were discovered in the past few decades stressing on their importance as regulators of cellular development and differentiation (Amaral et al., 2008). MicroRNAs, piwi-associated RNAs and endogenus small interfering RNAs are ncRNAs which held the attention of the scientific community for a long time (Castel and Martienssen, 2013; Yates et al., 2013). Nevertheless, in the last few years, the long non-coding RNAs (lncRNAs) have stood out to be the most prominent class of ncRNAs, defined as non-coding transcripts longer than 200 nucleotides (Cabili et al., 2011; Derrien et al., 2012; Guttman et al., 2009; Pauli et al., 2011a; Ulitsky et al., 2011; Young et al., 2012). They are not a recent addition to the ncRNA repertoire, with well characterised members reported in the past decades like the *Xist* (Brockdorff et al., 1992), *MALAT1 (Ji et al., 2003)* and *HOTAIR* (Rinn et al., 2007). A database cataloging experimentally verified lncRNAs from different organisms (lncRNAdb) (Amaral et al., 2011) reports 127 lncRNAs in human. Recent large-scale genome wide studies by the ENCODE consortium has reported ~10,000 lncRNA genes (Derrien et al., 2012; Djebali et al., 2012) in the human genome demonstrating that our current knowledge on lncRNAs is based upon 1% of the total estimated population. This indicates a requirement for more experimental validation than currently reported, to build a concrete hypothesis on the functionality of all the lncRNAs identified. It is expected that with sequencing technologies becoming more sensitive and cost-effective the number of predicted lncRNAs will increase, a recent study even

estimating (computationally) a total of 40,000-50,000 lncRNA genes in human and mouse genomes (Managadze et al., 2013). At this juncture the following aspects of lncRNAs biology need a better understanding:

- The cellular machineries exploiting their functions.
- Mechanisms of action and implication in organism development and disease.
- Conservation and significance with respect to evolution.
- Identification and validation.

1.2 The need for lncRNAs: Advantages over proteins and small non-coding RNAs

1.2.1 Long non-coding RNAs in the X inactivation centre

Suppression of the phenotypic effects of an additional X chromosome in the mammalian females, requires silencing of the gene expression in one X chromosome, also known as X Chromosome Inactivation (XCI). The X inactive specific transcript (*Xist*) lncRNA occupies the X inactivation center (Xic) and is reported to be involved in transcriptional repression throughout the X chromosome (Brockdorff et al., 1992; Brown et al., 1992; Clemson et al., 1996). A series of lncRNA based regulatory actions are performed at the X inactivation centre of one allele to initiate X chromosome inactivation (XCI) (**Figure 1.1**) (Lee, 2011). The evolution and mechanism of the X inactive specific transcript (*Xist*) is an ideal example to fathom the working intricacies of a lncRNA. The *Xist* lncRNA

binds to the Polycomb repressive complex 2 (*PRC2*) through a conserved repeat motif (*RepA*) and guides it to the Xi (inactivated X chromosome) (Zhao et al., 2008). The *PRC2* is an epigenetic complex which trimethylates histone H3 at Lys^{27} (H3K27me3) and facilitates stable maintenance of the X inactive state. The *Xist* transcript is aided in docking to the Xi by the *YY1*, which is a bivalent protein capable of binding both DNA and RNA (Jeon and Lee, 2011). Allelic control of *Xist* action is maintained by two other lncRNAs. The *Xist* antisense RNA (*Tsix*) represses transcription of *Xist* in one allele by mobilising a DNA methyltransferase (*Dnmt3a*) for *Xist* silencing (Sado et al., 2005) while the *Xist* activator *Jpx* positively regulates *Xist*, acting in trans and antagonistic to *Tsix* (Tian et al., 2010). A recent report shows the *Jpx* transcript interacting with the CCCTC-Binding Factor (*CTCF*) to evict it from the *Xist* locus resulting in *Xist* transcription (Sun et al., 2013b). The *Xist*, *Tsix* and the *Jpx* lncRNAs are transcribed from the Xic locus and are involved in the local chromatin remodeling. It is important to note that the Xic was not always non-coding in nature (Duret et al., 2006a), the shift concurring with the evolution of eutherian mammals 150 million years ago (Lee, 2009). This transition from coding to non-coding is not expected to be retained unless the presence of lncRNA proves to be advantageous in comparison to protein coding genes for upkeep of regulatory processes. Thus I will discuss below the distinct advantages of lncRNAs over protein-coding genes in performing regulatory functions in the cell.

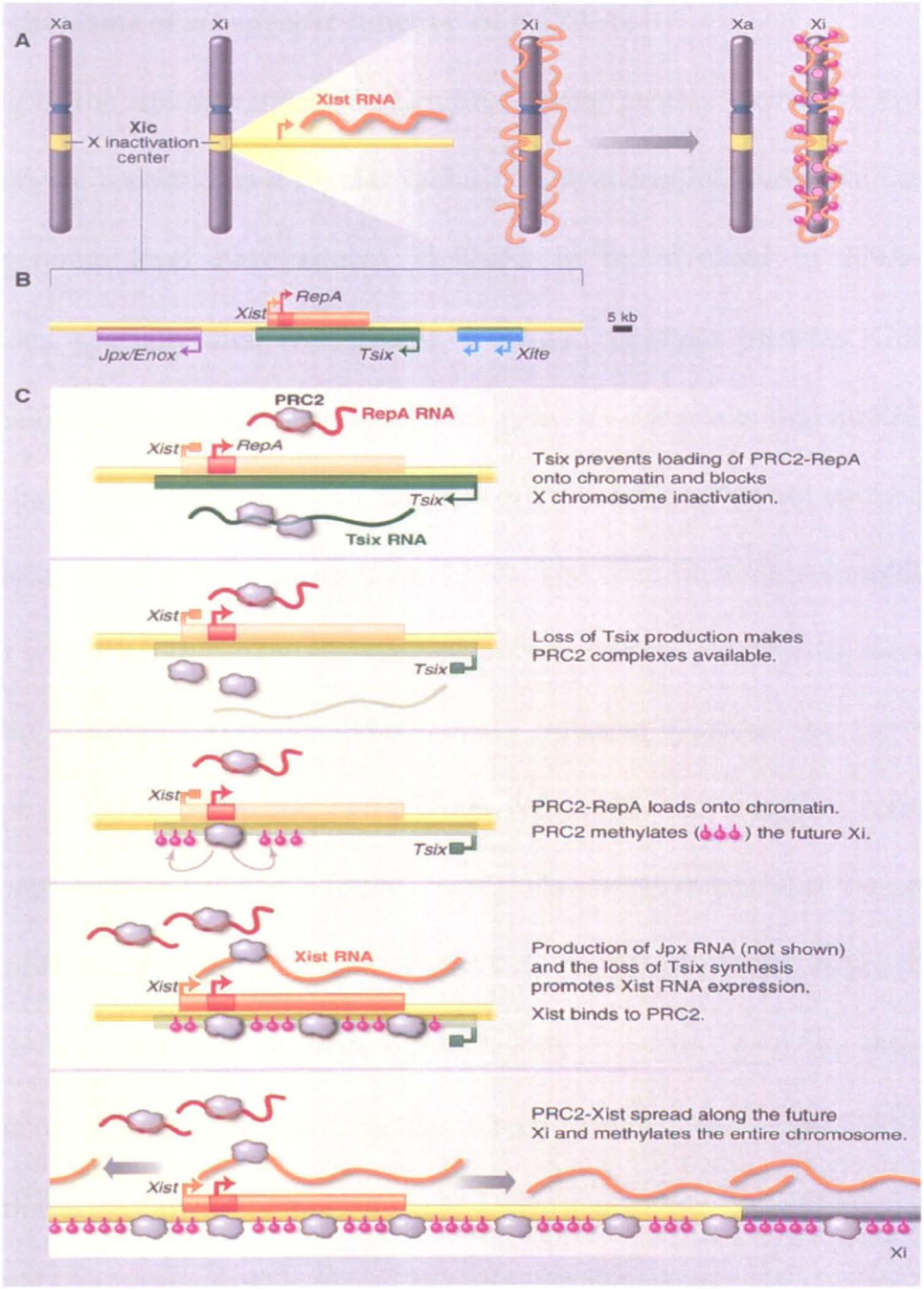


Figure 1.1 LncRNAs in X-chromosome inactivation. **A)** The lncRNA *Xist* is transcribed from the Xic of the inactive X chromosome. *Xist* RNA covers the entire chromosome and silences gene expression through epigenetic modification of histones and DNA. **B)** The core region of the Xic and its lncRNAs. **C)** LncRNA-protein interactions at the initiation of XCI. The figure is reproduced from Lee *et al*, 2011.

1.2.2 Mechanisms of *cis* specific function of lncRNAs

Tethering to the site of transcription and recruiting protein factors for epigenetic regulation are reported as a regular feature of cis-acting lncRNAs. Studies at the whole genome level have shown lncRNAs to be involved in RNA-protein interactions, specially their recruitment of the *PRC* complex proteins (Guil et al., 2012; Khalil et al., 2009; Zhao et al., 2010). A prior review posits that lncRNAs may utilise a bare 5' end for chromatin complex capture while allowing an incipient 3' to act as an anchor to a locus (Lee, 2009). The Xist lncRNA exemplifies this behavior where it induces the silencing of its chromosome of origin by recruiting a chromatin modifying complex (Wutz, 2011). Another lncRNA the *HoxA* distal transcript antisense RNA (*HOTTIP*) binds with a histone modifier complex to bring about histone Histone 3 lysine^4 trimethylation (H3K4me3) at the promoter regions of flanking coding genes in the *HoxA* cluster in human fibroblasts (Wang et al., 2011). The lncRNA *COLDAIR* is transcribed in plants from the intron of a coding gene (*FLC*) and in turn changes the epigenetic state in the *FLC* locus to control flowering time (Heo and Sung, 2011). In comparison to lncRNAs a protein-coding RNA molecule forfeits allelic and location awareness after transport to the cytoplasm for translation, while small non-coding RNAs are not of the ideal length to act as tethers. Another important aspect which assists the site specific action of lncRNAs is their stability. Half life of lncRNAs are on an average lower than that of protein-coding mRNAs and they exist in low copy-numbers (Cabili et al., 2011; Clark et al., 2012; Djebali et al., 2012). A quick degradation of the RNA molecule can be a limiting factor to its half-life, avoiding its displacement to other locales in

the cell. The *Tsix* transcript has a half-life of 30-60 minutes, its quick degradation allowing optimal RNA concentrations to reach only at the site of action leading to a strict *cis* mechanism (Sun et al., 2006).

1.2.3 Mechanisms of *trans* specific function of lncRNAs

It is not that lncRNAs exert themselves always in *cis* mode. There are *trans*-acting lncRNAs which disseminate from their locus of origin and act at large distances including other chromosomes except for particular cases where a purely *cis* or *trans* mechanism cannot be distinguished. An example being the *Xist*, which remains functionally active, inducing repressive chromatin state even when introduced as a transgene (Jeon and Lee, 2011). Trans-acting lncRNAs rarely tether to protein complexes for localised action or are involved in target site binding (Kornienko et al., 2013; Lee, 2012), an exception being a promoter associated ncRNA which meshes with the target site of the transcription factor *TTF-1* and in turn is recognised by a DNA methyltransferase in mouse fibroblasts (Schmitz et al., 2010). On the contrary they usually act as molecular scaffolds or co-activators and co-repressors. A prime example of scaffolding is the *Hox* transcript antisense RNA (*HOTAIR*) lncRNA which originates from the *Hoxc* locus (Rinn et al., 2007) and scaffolds *PRC2* and Lysine (K)-specific demethylase 1A (*LSD1*) proteins to alter the chromatin state of the *Hoxd* genes (Tsai et al., 2010). Examples of other mechanisms include the Steroid receptor RNA activator 9 (*SRA*), a lncRNA which can activate steroid receptor-dependent gene expression by binding with the nuclear receptor co-activators in human (Lanz et al., 1999) and a human Alu RNA, which interacts

with RNA Pol II complex at the promoter of target genes to repress gene expression during cellular heat shock response (Mariner et al., 2008). The transcription factors Octamer-Binding 4 (*Oct4*) and *Nanog* are reported to bind two lncRNAs to control the pluripotent state in mouse embryonic stem cells, these lncRNAs are not only governed by the transcription factors but they themselves act as co-activators to regulate the developmental state (Sheik Mohamed et al., 2010). In principle trans-acting lncRNAs may behave more like small RNAs or transcription factors, permeating large gene networks through initiation of a signaling cascade but bear an advantage of sequence space over other proteins and small ncRNAs, providing specific scaffolding and binding mechanisms for gene regulation. The sequence length of lncRNAs may aid in the formation of secondary structures giving binding specificity to particular protein complexes. Thus the eukaryotic cell may be imagined containing a diaspora of transcripts and proteins, with the concerted action of transcription factors and lncRNAs followed by downstream epigenetic programming. This results in multiple network specific combination of transcription factors, lncRNAs and epigenetic complexes to attain specific as well as global responses to different stimuli. Thus, the long non-coding RNAs are able to interact with chromatin modifying complexes, transcription factors as well DNA elements to regulate the expression of various genes in a highly specific manner. This diversity of function gives them a distinct advantage over coding genes and small RNAs to act as a regulatory molecule.

1.3 Functional diversity of lncRNAs

Long non-coding RNAs are implicated in diverse molecular mechanisms. The diversity comes from ability of the lncRNAs to interact specifically with protein complexes as discussed before. Currently it is almost each fortnight a novel lncRNA is being reported. It would not be an exaggeration to say that lncRNAs may soon rival proteins in the range of functions they perform. In this regard it is important to discuss in-depth the major mechanistic traits exhibited by lncRNAs due to their specific expression, length and stability.

1.3.1 Interaction with transcription factors

Long non-coding RNAs are reported to entice transcription factors away from their targets or vie for their DNA-binding sites during stress response and growth. The transcription factor Nuclear Transcription factor *Y (NF-Y)* has three subunits *NF-YA, NF-YB* and *NF-YC* (Manni et al., 2008). The *NF-Y* acts as a repressor (Ceribelli et al., 2006) and as a co-activator (Morachis et al., 2010) for multiple targets of the *p53* gene which are involved in apoptosis. A lncRNA, *P21* Associated NcRNA DNA damage Activated (*PANDA*) sequesters *NF-YA* away from *NF-YA/p53* co-regulated promoters during DNA damage response to evade apoptotic cell death (Hung et al., 2011). The growth arrest specific 5 (*GAS5*) lncRNA docks with the DNA-binding domain of Glucocorticoid Receptor (*GR*) and competes with glucocorticoid receptor elements for binding to the GR thus regulating the cellular metabolism during cellular growth arrest (Kino et al., 2010). These examples show

the ability of lncRNAs to play an active role in the molecular cross-talk between transcription factor and their target genes, providing an additional dimension for regulation of specific genes and pathways.

1.3.2 Regulation of nuclear compartment and splicing

Long non-coding RNAs are noted to be involved in nuclear compartment regulation. The nuclear compartment comprises of the nuclear bodies which are sub-nuclear organelles functioning in response to different cellular and environmental cues (Mao et al., 2011). The Nuclear Paraspeckle Assembly Transcript 1 (*NEAT1*) and Metastasis Associated Lung Adenocarcinoma Transcript 1 (*MALAT1*) are two candidate lncRNAs with well documented role in functioning of the nuclear compartment. The *NEAT1* helps in the formation and stability of nuclear paraspeckles in human cell lines (Clemson et al., 2009), the paraspeckles themselves are involved in nuclear retention of mRNAs (Chen and Carmichael, 2009). In contrast the *MALAT1* confines splicing factors to nuclear paraspeckles for phosphorylation (Bernard et al., 2010) and helps to localise the splicing factors to sites of transcription (Tripathi et al., 2010) regulating alternative splicing of mRNA precursors. Another splicing regulating lncRNA, the Myocardial Infarction Associated Transcript (*MIAT/Gomafu*) shows a restricted expression in mouse neurons and has its operation space curbed to the nuclear compartment of the cell (Sone et al., 2007). It has a conserved tandem repeat sequence which aids in its binding with Splicing Factor 1 (*SF1*), thus proposed to be involved in regulation of splicing efficiency (Tsuiji et al., 2011). A recent report found a marked correlation

between alternative splicing events and organism complexity with maximum splicing events observed in primates (Barbosa-Morais et al., 2012). The role of lncRNAs in regulating splicing events genome-wide might well prove to be an integral part of the developmental programming of an organism. However, it is yet inexplicable as to why the knockouts of neither *NEAT1* and *MALAT1* yield potent phenotypes (Eißmann et al., 2012; Nakagawa et al., 2011) suggesting that conventional knockout/knockdown studies correlating quantity to functionality might not hold strong for lncRNAs.

1.3.3 Post-transcriptional modifications and translational regulation

There are multiple examples of lncRNAs involved in post-transcriptional regulation of mRNA molecules. An antisense transcript coming from the 3'-UTR of inducible nitric oxide synthase (*iNOS*) interacts with its sense transcript as well as with AU-rich element-binding protein *HuR* to increase the stability of the *iNOS* in rat hepatocytes (Matsui et al., 2008). The *iNOS* protein positively regulates the secretion of nitric oxide (Nathan and Xie, 1994), and nitric oxide is implicated in a diverse range of cellular processes from angiogenesis (Fraisl, 2013), myogenesis (De Palma and Clementi, 2012) to programming of cellular differentiation pathways (Mujoo et al., 2011). In contrast to increasing stability, lncRNAs also aid in degradation of 1% mRNAs in human HeLa cell line as evident from the *Staufen1* mediated decay pathway. The *Staufen1* is a protein involved in degradation of translationally active mRNAs, which relies on imperfect base pairing between ALU elements on 3'UTR of a coding mRNA and lncRNAs called *half-STAU1-binding site*

RNAs for target site identification (Gong and Maquat, 2011). While lncRNAs can regulate the stabilisation or degradation of mRNA molecules they are also reported to regulate the level of protein expression, leaving the mRNA concentration undisturbed. The expression of a transcription factor (*PU.1*, regulating hematogenesis), is negatively affected at the protein level in human and mouse cell lines when its antisense non-coding transcript competes with it for binding to the eukaryotic translation Initiation Factor 4A1 (*eIF4A*) (Ebralidze et al., 2008). A recent report showed that a B2 SINE embedded in a mouse lncRNA antisense to Ubiquitin carboxyl-terminal esterase L1 (*Uchl1*) enhances the translation of the *Uchl1* transcript (Carrieri et al., 2012). The *Uchl1* gene is involved in brain specific protein degradation and is implicated in Parkinsons disease (Liu et al., 2002). This observation brings forth a new aspect of lncRNA mechanism involving close association with overlapping transposable elements. In fact another study reports the presence of 361 lncRNAs in mouse with B2 SINE elements suggesting that the previous report may not be an exclusive event (Kapusta et al., 2013). A lncRNA *lincRNA-P-p21* associates with RNA-binding protein human antigen R (*HuR*), leading to lower *lincRNA-p21* stability in a human cell line. During low *HuR* levels, *lincRNA-p21* is expressed and lowers the translation product of *JunB* and Catenin (cadherin-associated protein) Beta 1 (*CTNNB1*) coding genes in collaboration with the transcriptional repressor DEAD box helicase (*Rck*) thus eliciting a complex feedback loop (Yoon et al., 2012).

1.3.4 Cross-talk with small non-coding RNAs

An interesting aspect of lncRNA mechanism is its cross-talk with small ncRNAs, where lncRNAs compete, act as decoy or are the source for small ncRNAs. An antisense lncRNA originating from the β-secretase-1 (*BACE1*) locus augments the stability of the *BACE1* coding gene by masking its miRNA binding sites (for *miR-485-5p*) in human HEK293T cells (Faghihi et al., 2010). The *BACE1* is reported to be a critical gene involved in Alzheimers disease due to its involvement in beta-amyloid (abeta) peptide secretion (Vassar et al., 2009). Thus indirectly the *BACE1* antisense transcript plays a major role in maintaining the stability of *BACE1* mRNA and accumulation of abeta peptides in the brain. There are also examples of lncRNAs acting as decoys or sponges for miRNAs, having miRNA binding sites in their 3UTRs. A long non-coding RNA *HULC* (Highly Upregulated in Liver Cancer) acts as an endogenous sponge for the miRNA, *miR-372*, hence setting up a self regulatory loop mediated by the target gene of the miRNA (*CREB* phosphorylating protein) and the *CREB* protein which binds to the core promoter of the lncRNA in human tumorous liver tissue (Wang et al., 2010). The *HULC* sets an interesting precedent for understanding how an mRNA/RNA molecule may regulate a miRNA, since in the past the sole focus has been on how miRNA molecules regulate mRNA. An important finding in support of the endogenous sponge mechanism came from the lncRNA Muscle Differentiation 1 (*lincRNA-MD1*) which provides alternative binding sites for miRNAs *miR-133* and *miR-135* in human myoblasts (**Figure 1.2**) (Cesana et al., 2011). The miRNAs *miR-133* and *miR-135* control the expression of Mastermind-like 1 (*Maml1*) and Myocyte Enhancer Factor

2C (*MEF2C*) respectively. The *MEF2C* is reported to be an important transcription factor involved in muscle and cardiovascular development (Lin et al., 1997) and the *MAML1* is reported to regulate the transcription of *MEF2C* (Shen et al., 2006). Thus the *lincRNA-MD1* plays a role in muscle development and differentiation on the merit of its ability to block the activity of the miRNAs (*miR-133* and *miR-135*) by competing with the *MEF2C* and *Maml1* to bind with the miRNAs *miR-133* and *miR-135*. This behavior of lncRNAs is proposed as a part of a large regulatory network where the messenger RNAs and long non-coding RNAs cross-talk with the miRNAs as an intermediary component leading to an increase in the number of feasible signaling cascades (Salmena et al., 2011; Seitz, 2009).

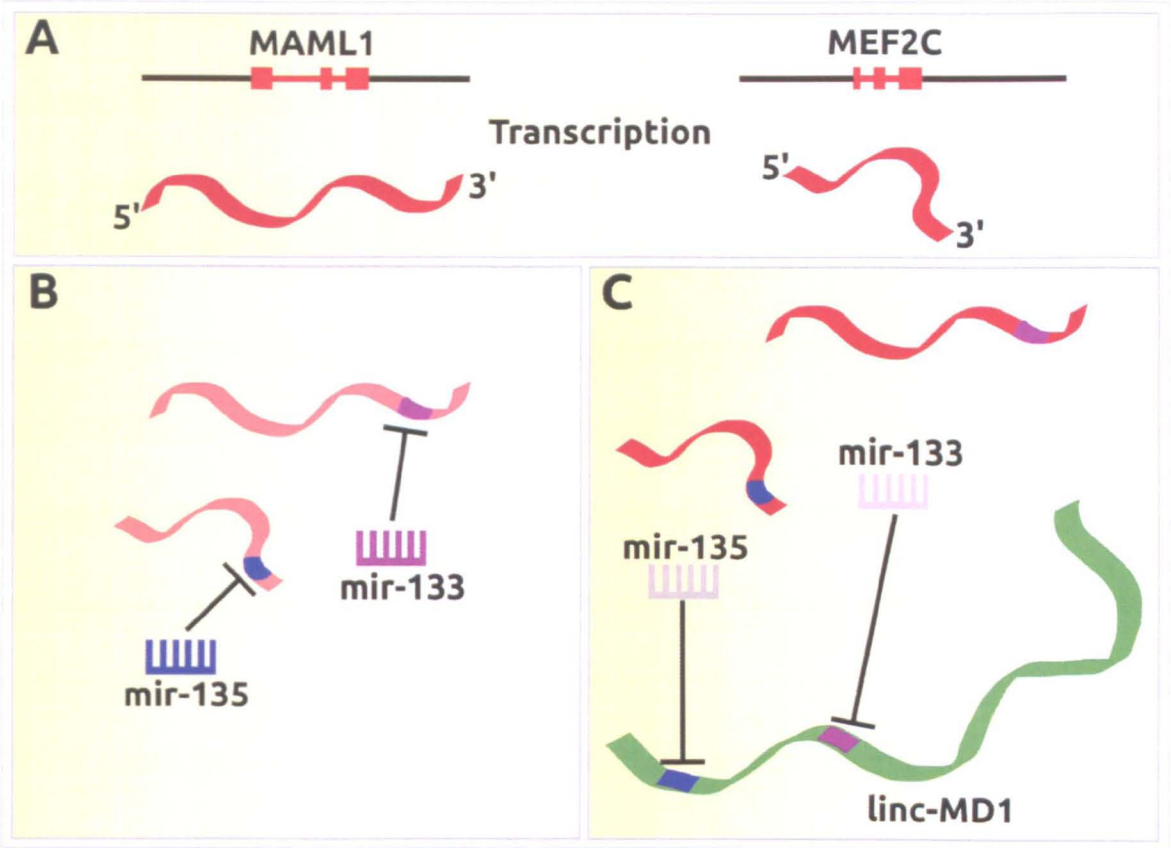


Figure 1.2 The role of *lincRNA-MD1* in regulation of genes important for muscle development and differentiation. **A)** Transcription of the *Maml1* and *MEF2C* genes. **B)** The normal scenario where the miRNAs (*miR-133, miR-135*) block the action of the *Maml1* and *MEF2C* genes. **C)** The alternative scenario where the *lincRNA-MD1* competes for the miRNA binding sites with *Maml1* and *MEF2C* hence preventing the miRNAs to block the translation of the coding genes. The mechanism depicted is described in Cesana *et al*, 2011.

Apart from competing with small RNAs, lncRNAs are also known to play host to small RNA transcription. The *H19* is one of the first lncRNAs reported to be involved in the imprinting of the coding gene Insulin-like Growth Factor 2 (*IGF2*) in mouse (Brannan et al., 1990). The *H19* gene has a microRNA (*miR-675*) embedded in its first exon, which targets the *IGF2* gene to control cellular growth (Keniry et al., 2012). The *Gas5* lncRNA gives rise to highly conserved snoRNAs

(Smith and Steitz, 1998) while *Gtl2*, *anti-Rtl1* and *Mirg* lncRNAs harbour multiple miRNAs and snoRNAs (da Rocha et al., 2008). Interestingly complementary base pair regions of sense/antisense transcript pairs, consisting of mRNAs and pseudogenes are reported to give rise to endo-siRNAs suggesting another less understood mechanism by which lncRNAs may exert their function (Tam et al., 2008; Watanabe et al., 2008). In fact a study associates the repression of *Xist* lncRNA by *Tsix* to a RNA-i mediated pathway, where *Dicer* dependent small RNAs originate from the complementary base pairing between the two lncRNAs (Ogawa et al., 2008). An important investigation on lncRNAs competing and hosting small RNAs could be whether it is the small RNAs which are the master switches regulating both lncRNAs and coding mRNAs or the lncRNAs themselves. Which means a better understanding of the knockdown/knockout/overexpression of the lncRNAs, to know whether the phenotype results from the lncRNA or its miRNA counterpart.

1.3.5 Long non-coding RNAs as enhancers

Enhancers are reported to be widely transcribed in mouse neuronal cells, giving rise to non-polyadenylated, non-coding transcripts, their expression levels correlating with that of nearby coding genes (Kim et al., 2010). Further a major fraction of mammalian RNA pol II initiation events in intergenic regions are reported to be associated with enhancers (De Santa et al., 2010). A previous study had shown the ability of lncRNAs to function as enhancers, inducing the expression of their neighboring coding genes but in a RNA-dependent fashion

(Ørom et al., 2010a). A recent report demonstrated the presence of enhancers within introns of coding genes, which give rise to stable lncRNA transcripts (Kowalczyk et al., 2012). These lncRNAs were called enhancer RNAs (eRNAs) as they worked in a manner similar to classical enhancers. However it is not yet well understood, whether the transcribed enhancers and lncRNAs acting as enhancers are part of similar or different regulatory pathways. However, a recent study demonstrated an enhancer-like mechanism of an lncRNA involving the *Mediator* complex. Amongst metazoans the *Mediator* multiprotein complex is reported to be vital in regulation of a diverse set of protein coding genes (Malik and Roeder, 2010). The *Mediator* complex is comprised of multiple subunits which are conserved across evolution and interact with various regulatory molecules like transcription factors, coactivators and repressors to regulate the expression of various genes. A class of ncRNAs called the ncRNA-activating (*ncRNA-a*) were reported which function in an enhancer like manner by binding with the subunit of the *Mediator* protein complex (*Med12*), and activate their neighboring genes in *cis* (Lai et al., 2013). Interestingly chromosome conformation capture assays showed chromatin looping between the *ncRNA-a* and their target genes, the looping reduced on depletion of either the *ncRNA-a* or the mediator subunit. This finding provides a valuable insight into possibly one of the principal mechanisms employed by enhancer like lncRNAs or eRNAs to activate the transcription of their neighboring genes.

1.4 Long non-coding RNAs in development and disease

A deeply intertwined circuitry of lncRNAs and transcription factors help maintain the pluripotent state of the cell (Guttman et al., 2011). This close association with the cell fate gives credence to lncRNAs as a major player in the developmental programming of the cell and suggests possible association with multiple diseases. Actually 43% of disease associated SNPs in human are known to lie in intergenic regions as compared to 45% in introns of coding genes, the rest present in exons or UTRs of coding genes (Hindorff et al., 2009). Numerous reports of lncRNAs with respect to metabolical, genetical and developmental disorders led to the creation of a database (LncRNADisease) exclusively for lncRNA disease associations (Chen et al., 2013a). The database holds information of ~500 lncRNAs which have experimental support to be involved in a particular disease. A computational method, the Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) was published recently to associate lncRNAs with probable diseases based on their expression profile (Chen and Yan, 2013).

1.4.1 Long non-coding RNAs in cancer

Cancer of any form or type closely correlates with an altered developmental programming of the cell. There are numerous reports implicating long non-coding RNAs in cancer (Cheetham et al., 2013). The *MALAT1* and the *HOTAIR* have also been well characterised for their implication in cancer. The *MALAT1* interacts with splicing factors to regulate alternative splicing of mRNAs in the nuclear compartment of the cell during cell division (Tripathi et al., 2010). Apart from

being a splicing regulator the *MALAT1* governs the gene expression of multiple target genes exclusively in lung carcinoma (Gutschner et al., 2013). Further it is reported to be overexpressed in multiple tumour types and associated with patient survival, indicating it to play a major role in cancer metastasis (Schmidt et al., 2011). However another recent study proved that the gene regulatory feature of *MALAT1* is not limited to cancerous cells, since it is linked to regulation of cell cycle genes required for G1/S and mitotic progression in normal human fibroblasts (Tripathi et al., 2013). The *HOTAIR* lncRNA acts in *trans* by recruiting the *PRC2* complex to silence the expression of genes in the *Hoxd* locus (Rinn et al., 2007). The *Hox* cluster genes are master regulators of embryonic cell development and differentiation, whose mis-regulation leads to human disease especially cancer (Barber and Rastegar, 2010). The *HOTAIR* transcript is reported to be involved as a proto-oncogenic factor in pancreatic, colorectal, hepatocellular and gastrointestinal cancer (Geng et al., 2011; Kim et al., 2013b; Kogo et al., 2011; Niinuma et al., 2012). Thus the mechanisms of *MALAT1* and *HOTAIR* prove to be critical paradigms to understand the concerted functions of lncRNAs in development and disease in vivo.

1.4.2 Long non-coding RNAs in neuronal disease

A functional aspect to which lncRNAs are often related is the development of the brain or neural tissues. A recent study has identified lncRNAs specific to mouse neural stem cells with a potential role in neurogenesis (Ramos et al., 2013). Previously lncRNAs transcribed specifically in the mouse brain were identified

from in-situ hybridization data, proposing a complex expression interplay with proximal coding genes, which are mainly of neurological importance (Mercer et al., 2008) followed by another report showing dynamic lncRNA expression pattern during neuronal and glial cell differentiation (Mercer et al., 2010). Two other studies report the presence of novel lncRNA transcripts dynamically regulated during development in human (Lipovich et al., 2013) and rat cerebral cortex (Wood et al., 2012). A subset of lncRNAs specific to mouse central nervous system (Ponjavic et al., 2009) and human retinal neurons (Mustafi et al., 2013) show constraint of sequence amongst mammals. This suggests a small group of lncRNAs catering to core neural developmental functions while the rest arise from a lineage specific evolution. A number of lncRNAs like the *HOTAIR* and *CRNDE* were found to be differentially expressed during the transition of embryonic pluripotent cells to neurons indicating their importance in differentiation and neuropsychiatric diseases (Lin et al., 2011a). The β-secretase enzyme, beta-site APP cleaving enzyme-1 (*BACE1*) is functionally important during synaptic transmission and myelination in the brain (Vassar et al., 2009). It is implicated in the Alzheimers disease due to its involvement in formation of amyloid beta (Aβ) in diseased brains (Kandalepas and Vassar, 2012). The BACE1 antisense transcript (*BACE1-AS*) masks miRNA binding sites (for *miR-485-5p*) in *BACE1* to increase the *BACE1* mRNA stability (Faghihi et al., 2010) and the antisense transcript shows elevated levels of expression in patients with Alzheimers disease (Faghihi et al., 2008). The nuclear enriched abundant transcript 1 (*NEAT1*) is an important constituent for nuclear paraspeckle formation (Clemson et al., 2009). The *NEAT1* transcript is reported to

interact directly with TAR DNA-binding Protein-43 (*TDP-43*) and fused in sarcoma/translocated in liposarcoma (*FUS/TLS*) in amyotrophic lateral sclerosis (ALS) motor neurons (Nishimoto et al., 2013). The *TDP-43* and *FUS* proteins are implicated in the ALS diseased state. This report adds another functionality to the *NEAT1* transcript where it may act as a scaffold for RNA binding proteins in the nuclei of ALS motor neurons. A lncRNA *CRNDE* (*Colorectal Neoplasia Differentially Expressed*) is highly expressed in multiple cancer cell types, shows a prominent expression pattern in human and mouse brain and promotes neuronal differentiation (Ellis et al., 2012). The *HOTAIR, MALAT1, CRNDE* examples suggests that a single lncRNA may provide a regulatory stimulus to different pathways implicated in cellular development, differentiation or the onset of a diseased state.

1.4.3 Potential of lncRNAs as therapeutic agents

The traditional drug targets of the genome, the coding genes, represent a small fraction of the genome (Overington et al., 2006) and the microRNAs inhibit expression at the translation level and are not known to be highly locus specific (Lim et al., 2005). In comparison lncRNAs, specially those which act in *cis* are known to activate or repress transcription in a locus specific manner and their inhibition can lead to a natural up-regulation of their target coding genes in contrast to traditional enzyme replacement therapies. A recent review suggests that the lncRNAs (specifically natural antisense transcripts, *NATs*) are an ideal molecule to fill the dearth of therapeutic targets for human diseases since they are

locus specific cis-regulators of their target coding genes (Wahlestedt, 2013). The *NATs* are transcribed opposite to the sense strand of protein-coding genes which results in them partially overlapping the exons, promoter and regulatory binding sites of the protein-coding gene (Faghihi and Wahlestedt, 2009). The *NAT* expression can be inhibited by single stranded oligonucleotides as exemplified by the seven fold rise in expression of the Brain-Derived Neurotrophic Factor (*BDNF*) on inhibiting its antisense counterpart (Modarresi et al., 2012). The lncRNAs are hitherto an untapped potential, most of them being novel transcripts with unknown function. Even at low expression levels lncRNAs can direct a specific regulatory mechanism, hence may require a smaller dosage of inhibitory oligonucleotides, reducing toxicity and off-target effects thus making them ideal candidate for therapeutic purposes. The basic mechanism of action, known function and organism of origin of the lncRNAs discussed so far are summarised in **Table 1.1**.

Name	Mechanism	Function	Class	Organism
Xist	Tethering chromatin modifying complex	X chromosome inactivation	cis-acting	Human, mouse
Tsix	Tethering chromatin modifying complex	*Xist* inactivation	cis-acting	Human, mouse
Jpx	Eviction of *CTCF* bound to the *Xist* locus	*Xist* activation	trans-acting	Human, mouse
MALAT1	Binding with splicing factors and transcription factors	Cell cycle control, cancer metastasis	trans-acting	Human, Mouse
Gomafu	Binding with splicing factor	Regulation of splicing efficiency in the nucleus	trans-acting	Mouse, Chicken
NEAT1	Formation of nuclear paraspeckles	mRNA export from nucleus to cytoplasm	trans-acting	Human

HOTAIR	Molecular scaffolding of chromatin modifying complexes	Organism development	trans-acting	Human
CRNDE	Scaffolding of chromatin modifying complexes	Insulin signalling, Cancer, Brain development	trans-acting	Human
HOTTIP	Tethering chromatin modifying complex	Organism development and angiogenesis	cis-acting	Human
COLDAIR	Tethering chromatin modifying complex	Control of flowering time	cis-acting	*Arabidopsis thaliana*
lincRNA-TTF1	Competing with transcription factor binding site	Ribosomal RNA regulation	trans-acting	Mouse
SRA	Binding with transcriptional coactivator	Development and reproduction	trans-acting	Human
SINE B2 RNA	Transcriptional repression through RNA POL II binding	Heat shock response	trans-acting	Human
PANDA	Sequestering of transcription factor	DNA damage response	trans-acting	Human
GAS5	Competing with DNA binding site of glucocorticoid receptor	Regulation of metabolism during cell growth arrest	trans-acting	Human
iNOS-AS	Interaction with sense RNA to increase its stability	Angiogenesis, Myogenesis	cis-acting	Rat
BACE1-AS	Stability of sense mRNA by protection against miRNA binding	Alzheimers disease	cis-acting	Human
Lnc-STAU1	Transactivation of RNA degrading protein	RNA binding mediated decay	trans-acting	Human
PU.1-AS	Repression of sense mRNA translation	Haematopoiesis	cis-acting	Human, mouse
LincRNA-UCHL1	Activation of sense mRNA translation	Brain development, neurodegenerative disease	cis-acting	Mouse
LincRNA-p21	Regulation of translation of multiple target genes	Regulation of cellular translation machinery	trans-acting	Human
HULC	Competing with miRNA target for miRNA binding	Cell proliferation, Tumour metastasis	trans-acting	Human
LincRNA-MD1	Competing with miRNA target for miRNA binding	Myogenesis	trans-acting	Human
H19	Imprinting	Insulin metabolism, cell proliferation	cis-acting	Mouse

Table 1.1 Summary of lncRNAs with known mechanism of action and function.

1.5 Evolution and conservation of lncRNAs

1.5.1 Conservation of sequence in the non-coding genome

The base pretext in case of a conservation analysis is that the sequence similarity between organisms of diverse taxa suggests its functional significance. In case of protein-coding genes the pressure to retain their amino acid composition proves to be a major factor in conservation of nucleotide sequence. Long non-coding RNAs are not bound by such a constraint and their large sequence space can possibly allow transcripts with no sequence conservation to perform similar functions on merit of the mRNA folding and vital pockets of binding sites. An understanding of the organisational change brought by species diversification in the non-coding regions is critical to gain insight into evolution and conservation of lncRNAs. Mouse non-coding sequences are reported to be conserved no better than a evolutionary neutral model, indicating them to be either non-functional or species specific (Wang et al., 2004). That non-coding regions lack conservation of sequence is not a universal truth considering the study mentioned above chose to omit 5% of the highly conserved non-coding part of the mouse genome. Conserved blocks of non genic regions between human and mouse were first identified as non-coding regions with a potential regulatory function (Dermitzakis et al., 2002). Further, genomic regions with perfect sequence identity between human, mouse and rat were reported as ultra conserved elements (UCEs) and contemplated to play a role in the ontogeny of organisms (Bejerano et al., 2004). Not all UCEs reported are non-coding in nature, some overlap exons and introns of coding genes. Conserved non-coding elements (CNEs) between human and fish were identified later

(Woolfe et al., 2005) along with ultra conserved non-coding genomic regions (UCRs) between human and mouse (Sandelin et al., 2004). Their high sequence identity set a precedent for their functional importance, for the sake of simplicity all such elements are addressed together as CNEs. A crucial observation from these studies was the clustering of CNEs around coding genes which have similar functions, particularly regulation of transcription and cellular development. There were already reports of short conserved non-coding enhancers regulating the expression of the *Sox9* and the *Hoxd* gene clusters (Bagheri-Fam et al., 2001; Santini et al., 2003) prior to the detection of the CNEs. Thus was conjectured that the CNEs cluster near their presumptive targets, that is, coding genes important in early development (especially DNA binding proteins), and play an active role in their regulation. The position of CNEs as potential enhancers was cemented by a study showing 45% of such elements act as tissue specific enhancers during early development, a majority directing the development of the nervous system (Pennacchio et al., 2006). Genomic regions containing arrays of conserved elements between mammals and teleost fishes encompass CNEs and their target developmental/transcription factor (*trans-dev*) genes along with functionally unrelated "*bystander*" genes (Akalin et al., 2009; Kikuta et al., 2007a). These regions are called genome regulatory blocks (GRBs) (**Figure 1.3**). The regulatory structure of GRBs was found to be replicated in insect genomes, along with extensive conservation of CNEs/target genes at the microsyntenic level categorizing them as an ancient regulatory feature of metazoan genomes (Engström et al., 2007). It is but a small percentage (< 20 %) of GRBs in human and insects which are identified as

ancestral associations, the rest originating in a lineage specific fashion (Irimia et al., 2012). The same study reported the widespread loss of GRBs, every bilaterian ancestral GRB being lost at least once in a metazoan while each metazoan species having lost at least a dozen conserved GRBs.

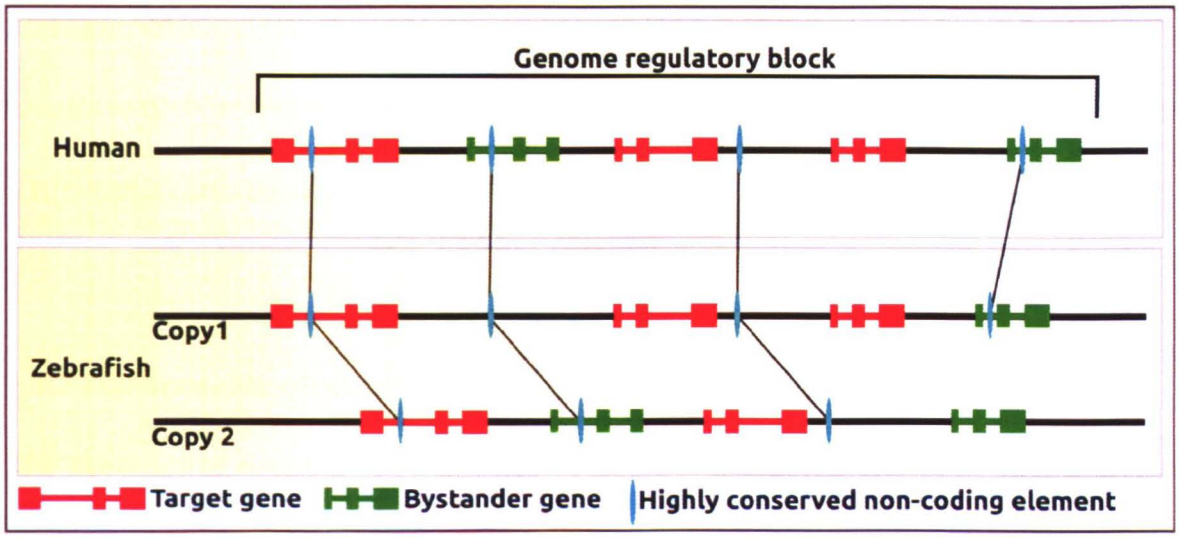


Figure 1.3 The definition of a genome regulatory block, its retention and distribution after a whole genome duplication event in teleost fishes. In the case of a HCNE acting as a regulator of a target gene, both the HCNE and the target gene are retained in duplicate copies in the zebrafish while a bystander gene may be lost. Conversely the HCNE may be lost along with its target gene in one copy. The structure of the figure is borrowed from Akalin *et al*, 2009.

CNEs identified in three different taxa (insects, worms and vertebrates) do not show sequence similarity, but remain associated with and regulate genes involved in highly similar functions, specially organism development (Vavouri et al., 2007). However, conserved non-coding sequences in a tunicate (*Ciona intestinalis*) are reported to share short segments of conservation with vertebrate CNEs (45 bp; 55% identity) and termed as oCNEs. These elements can act as enhancers in

transgenesis and cross-transgenesis experiments suggesting an evolutionary conservation of their cis-regulatory function. Interestingly the oCNEs are observed to be present in non-syntenic locations between vertebrates and urochordates, suggesting their co-option into novel regulatory networks due to chromosomal re-arrangements or retrotransposition (Sanges et al., 2013). Indeed, such elements results to also be transcribed and enriched in overlapping eRNAs suggesting again a functional link between noncoding transcription and regulatory functions.

1.5.2 Transposable elements and long non-coding RNAs

The presence of regulatory elements within lncRNAs genes can be an active driver of their evolution. A yet unexplored hypothesis in lncRNA evolution is the role played by overlapping enhancers or transposons. Not much was known about the effect of transposable elements (TEs) except for them playing an active role in the regulation and diversification of non-coding exons (van de Lagemaat et al., 2003; Zhang and Chasin, 2006). A recent study reported more than 80% of human lncRNAs containing TEs in sharp contrast to protein-coding genes (Kelley and Rinn, 2012). Further the TEs showed a positional bias for the transcriptional start site of lncRNAs, insinuating their possible role in regulation of the non-coding genes. Another study reported the presence of TEs in lncRNA exons of human (75%), mouse (68%) and zebrafish (66%) more than that expected by chance and the TEs were more likely to overlap a transcriptional start site or polydenylation site in a lncRNA as compared to a coding gene (Kapusta et al., 2013). An observation which catches the attention is the fact that TE sequences within

lncRNA exons are more conserved than random genomic regions or TEs in intronic regions or lncRNA exonic regions without TEs. Further TEs lying upstream of cell-type specific lncRNAs in human were found to possess an active chromatin state in the particular cell-type, leading to the inference of them working as cis-regulatory elements for lncRNAs (Kapusta et al., 2013). Promoters of very long intergenic RNAs (vlincRNAs) are recognised to overlap endogenous retroviral elements and the expression of vlincRNAs with such promoters correlates strongly with the level of malignancy of a normal cell type (Laurent et al., 2013). To summarise, TEs are found to be important features in modulating the sequence and expression of lncRNAs and potentially assigning functional constraints. Thus lncRNAs can be effectively classified on the basis of their location and overlap with other genomic features (**Figure 1.4**). Their genomic position can help in prediction of putative functional roles, considering that unlike protein-coding genes for lncRNAs the classical approach of relying upon sequence conservation to predict function is not effective.

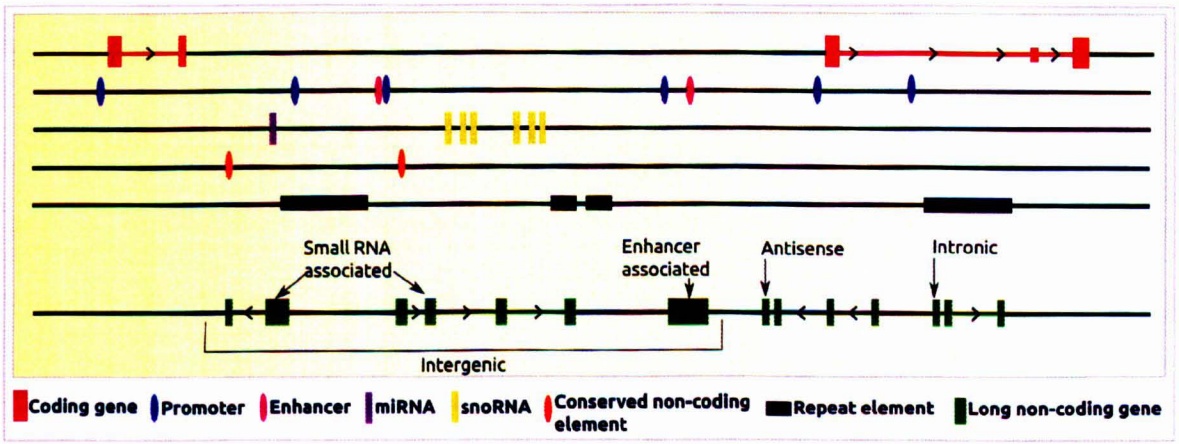


Figure 1.4 Classification of lncRNAs based on their genomic position and overlap with other non-coding RNAs, protein-coding genes and regulatory features.

1.5.3 Conservation of sequence in lncRNAs

Assessment of sequence constraint is defined as the proportion of the nucleotide substitution rate in functional sequence, which can be categorized into neutral, unconstrained, and constrained. Long non-coding RNAs show a lower sequence constraint in comparison to coding and small non-coding RNAs, the average nucleotide substitution rate for intergenic non-coding RNAs being 90-95%, implying 5-10% of sequence conservation (Ponjavic and Ponting, 2007). There are sporadic reports of sequence conservation in lncRNAs, primarily in case of the *Xist*, *Sox2ot*, *Har1F* and *HOTAIR* lncRNA genes. The *Xist* lncRNA, shows sequence conservation in 14 vertebrate species but has no homologs in non eutherian vertebrates (Duret et al., 2006a). The length of the *Xist* transcript has small pockets of sequence conservation, specially in its fourth exon and in five internal repeat element sequences (repA-E) (Pontier and Gribnau, 2011). The repA sequence is known to be important for *Xist* functioning, forming a hairpin structure to bind the *PRC2* protein complex (Zhao et al., 2008) while the other repeats and exon 4 functions are not well characterised. The size and structural orientation of *Xist* conserved regions are reported to be the possible factors guiding its localisation and X inactivation mechanism. The *Sox2* overlapping transcript (*Sox2ot*) and Human accelerated region 1F (*Har1F)* are two other lncRNAs which have vertebrate conserved sequence elements spanning the transcript (Amaral et al., 2009; Pollard et al., 2006a). The human *HOTAIR* lncRNA shows two conserved regions in comparison with the mouse genome but appears to evolve faster than the nearby *Hoxc* genes, yet shows a considerable conservation in secondary

structure (He et al., 2011a). An interesting observation on the human *HOTAIR* lncRNA is that, of its two functional regions conserved with mouse genome only one falls within the murine *HOTAIR* transcript and the deletion of the lncRNA in mouse has no visible effect on expression of the *Hoxd* genes (Schorderet and Duboule, 2011). This may either be an error because of incorrect annotation in mouse or more convincingly appears to be a case of rapid evolution of the gene to perform a vital function in primates. Forty three putative lncRNAs in chicken show conservation with human, rat and mouse transcripts at greater than 80% sequence identity (Hubbard et al., 2005). Around ~600 lncRNAs were identified showing constraint in their nucleotide substitution rates between mouse and human, those expressed in brain showing higher degree of sequence and secondary structure conservation (Ponjavic et al., 2009). In another similar study brain specific mouse lncRNAs in bird and opossum were reported to be highly variable at the sequence level but their putative promoter regions, exon–intron boundaries and the pattern of expression during embryonic and early postnatal stages show pronounced evolutionary conservation (Chodroff et al., 2010). Performing a stringent sequence homology search (< 0.05% false positives) of mouse lncRNAs against vertebrate conserved elements in the zebrafish genome gave me a small figure of 4-11% of sequence conservation corroborating the previous reports on lncRNA sequence identity (Basu et al., 2013). A recent study identified a small set of lncRNAs (~20) specifically expressed in human retinal neurons showing sequence conservation in mammals suggesting the existence of conserved lncRNA subsets functioning in retinal and visual maintenance of

mammals (Mustafi et al., 2013). In contrast to the lncRNA genes their promoters are reported to show sequence conservation at par with those of coding genes implying the need for a constrained transcription pattern (Carninci et al., 2005; Derrien et al., 2012; Guttman et al., 2009). The above reports of lncRNA sequence conservation suggest a rethink for consideration of primary sequence information as a perpetual measure for functional identity. Still more conclusive evidence maybe obtained only when making comparisons of more exhaustive lncRNA populations rather than relying upon genomic alignments or incomplete lncRNA catalogues. It may well be the conservation of splicing/expression pattern, secondary structure and genomic locus of origin which help in the *de-facto* retention of the lncRNA function.

1.5.4 Positional conservation of lncRNAs with respect to their flanking coding genes

There is very little information on the positional conservation of lncRNAs during evolution. For example, protein-coding genes lying near a lncRNA gene in zebrafish have a higher probability to have orthologs adjacent to lncRNA genes in human or mouse (Ulitsky et al., 2011). There is an obvious catch to this statement, which is the percentage of total lncRNAs sampled for making such an observation, envisaging many lncRNAs yet lie undetected resulting in a bias towards lncRNAs falling inside syntenic loci. Indeed a report estimates the total number of mammalian lncRNAs much beyond the numbers currently identified (Managadze et al., 2013). This study employed the use of conserved orthologous regions

between the mouse and human genomes to estimate around 2/3rd of lncRNAs to be orthologous in mammals. This estimation may be slightly far fetched since the authors decide to completely neglect factors like microsyntenic association and orientation with respect to proximal coding genes and rely upon whole genome alignments. Another study reported the presence of conserved microsynteny between coding genes separated by no more than four other coding genes across the metazoan lineage (Irimia et al., 2012). Such a study gives impetus to check for linkage of protein-coding genes and lncRNAs over large evolutionary distances. While singularly it is difficult to predict lncRNA conservation due to their lack of sequence homology, in principle the positional association with a protein-coding gene across multiple species may reflect a functional constraint which results in transcription of a lncRNA in the particular locus. The functional constraint may be either the coding and non-coding genes being guided by common regulatory programs or it may result in the identification of a lncRNA subgroup which retain their positional identity to perform a cis-regulatory function. Prediction of such a subgroup of lncRNAs will contribute immensely in understanding the putative functions and mechanisms which may drive the evolution of lncRNAs.

1.6 Strategies for identification of lncRNAs

1.6.1 Computational strategies for lncRNA identification

Filtering by the length of transcripts, their overlap with coding genes and synteny amongst mammals forms the basis of a lncRNA discovery pipeline (Khachane and

Harrison, 2010). An alternative strategy was employed in another study relying upon the size of ORF and homology to a protein family database for lncRNA detection (Jia et al., 2010). The length of the transcript (> 200 nucleotides), size of ORF (< 100 AA), lack of homology to an annotated protein and a low coding potential are the major parameters for computational prediction of lncRNA transcripts used recurrently (Cabili et al., 2011; Pauli et al., 2011a; Ulitsky et al., 2011). The potential of a transcript to code for a protein is an important aspect to identify a lncRNA. In the past, codon substitution frequency analyses were used as a reliable metric to gauge the coding potential of a transcript (Lin et al., 2008). It is simply meant to assess the conservation of amino acid codons across evolution as a measure of the coding potential of a transcript. This method was later published as the program PhyloCSF but was restricted to species with complete genome sequence due to its need for a multi species nucleotide alignment (Lin et al., 2011c). Alternative approaches considering the length of open reading frames (ORF) and sequence homology removed the need for whole genome sequence. The Coding Potential Calculator (CPC) program estimates the coding potential by comparing the size and integrity of a predicted ORF along with its sequence homology against a large protein family database (for ex. UniProt or NCBI NR) (Kong et al., 2007). An alternative but faster program (PORTRAIT) omitted the homology information and relies upon translation of sequences across all reading frames followed by a machine learning algorithm on the *ab-initio* properties of the longest ORF and the nucleotide composition to estimate the coding potential (Arrial et al., 2009). Another program CPAT, functioning on the lines of the PORTRAIT software

uses the metrics of ORF prediction to measure the coding potential and claims to outperform both CPC and the PhyloCSF in terms of prediction accuracy (Wang et al., 2013). The prediction of ORF brings us to the question of translation of small peptides embedded within long transcripts which may be predicted as non-coding due an oversight by the software programs. It is preliminary to appoint a conclusive view on the best method to classify a transcript as coding or non-coding, yet the computational measurement of coding potential appears to be a fast and cost-effective method to get a reasonable estimate of the ncRNA populace. Currently there is a lack of an easily implementable computational pipeline for lncRNA identification whilst there exist numerous published studies adopting a few core computational parameters to detect lncRNA transcripts (**Figure 1.5**).

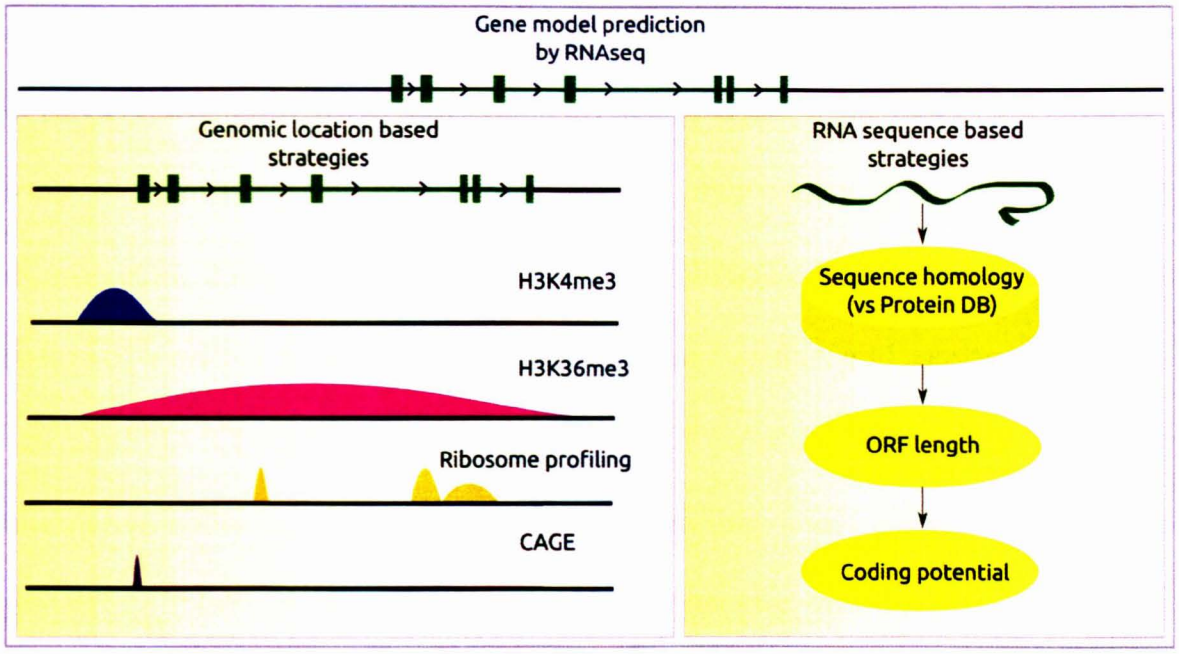


Figure 1.5 Major strategies defining computational prediction of lncRNAs. Genomic location based strategies involve validating the lncRNA structure by comparing it against H3K4me3 chromatin modification and CAGE peaks to identify the transcriptional start. Further the H3K36me3 modification peaks reflect on the length of the transcribed element and Ribosome occupancy reflects its potential to be a protein-coding gene. Sequence based strategies compare the lncRNA sequence against a public protein coding database followed by a filtering based on ORF size (< 100 amino acids) and measurement of coding potential.

1.6.2 Experimental strategies for identification of lncRNAs

Apart from the computational measures the single important factor which may define a lncRNA transcript is its inability to translate into a protein or a small peptide. The gold standard for measuring the translational potency of a lncRNA is an in-vitro translation experiment (Brannan et al., 1990; Brockdorff et al., 1992). However such an assay is not optimized for large scale studies and there is a possibility of getting a false negative result even in the case of known coding

genes. Recent evidence in the *Drosophila melanogaster* genome shows the presence of conserved small ORFs in transcribed intergenic regions (Ladoukakis et al., 2011), some of them lying within ncRNAs have the ability to code for short peptides (Kondo et al., 2010). A particular example is of SRA/*SRAP* which exhibits the dual nature of a ncRNA transcript, both as a regulatory ncRNA and as a translationally active mRNA (Chooniedass-Kothari et al., 2004; Kawashima et al., 2003), where the alternative splicing of an intron begets the coding and functional non-coding transcripts (Hube et al., 2006). Though this may not be a regular behavior in lncRNAs, dismissing the functionality of short peptides emanating from a lncRNA molecule would mean ignoring the "*pervasive translation*" in a genome. However, an argument favoring the ncRNAs is that the amino acid component and physiochemical properties of peptides arising from small ORFs can be so diverse from the known proteome that, without direct functional evidence, it is arduous to imagine such an alternative peptide universe (Cruveiller et al., 2007). Ribosomal profiling, an experimental method, determining putative RNAs bound to ribosomes, is an alternative approach to identify RNAs encoding small peptides (Pueyo and Couso, 2011). Ribosome-profiling in mouse ES cells revealed that the majority of lncRNAs are in fact associated to ribosomes (Ingolia et al., 2011). However, in the zebrafish genome numerous lncRNAs are observed to share a ribosome profile similar to 5' UTRs of coding genes or coding genes themselves (Chew et al., 2013) and an additional study reported that the majority of lncRNAs share similar ribosome occupation profiles as that of classical ncRNAs and 5UTRs and hence the occupancy metrics alone was not deemed sufficient to classify a

transcript as being translated (Guttman et al., 2013).

1.7 Large scale discovery of lncRNAs in metazoans

In the last five years large scale computational identification of lncRNAs in different cells, tissues and developmental stages of an organism has seen a steep rise in numbers. Such an analysis is dependent on technologies like microarrays, tiling arrays and RNAseq. RNAseq has now proven to be the method of choice for transcriptomic studies as it provides a direct quantification of the cDNA population, independent of the known fraction of transcriptome. Mapping of short reads on the genome and sequence assembly of the mapped reads into transcripts play a pivotal role in the accurate build of gene models specially lncRNAs from RNAseq data. Tophat, Bowtie, MapSplice and Star are the widely used programs for mapping of short reads on the genome (Dobin et al., 2013; Kim et al., 2013a; Langmead and Salzberg, 2012) while Cufflinks and Scripture are commonly used to assemble the mapped reads into transcripts (Guttman et al., 2010; Trapnell et al., 2010).

1.7.1 The Ensembl, FANTOM and ENCODE projects

Before discussing specific studies identifying lncRNAs in different organisms it is important to discuss the Encyclopedia of DNA Elements (ENCODE), the Functional Annotation of the Mouse (FANTOM) and the Ensembl, projects which stand as premier consortiums directed towards annotation of the human and

mouse genomes specially the non-coding portion. The ENCODE project starts where the Human Genome Project finished, with an aim towards identification of all *"functional"* DNA sequences that may be transcribed (both coding and non-coding) as well those which may play a regulatory role without getting involved in the act of transcription (Maher, 2012). The project was started with a pilot phase where 1% of the genome was annotated using various high throughput technologies resulting in the identification of a complex pattern of widespread transcription in the selected genomic regions (Birney et al., 2007). Additionally, numerous intergenic regions with a potential role in regulation of gene expression were also identified. The technologies and the methods which were standardised during the pilot phase, along with the incorporation of several new technologies, were further applied on the entire genome, to generate ~1600 datasets from 147 different cell types, the results of which were published as 30 scientific papers. The principal ENCODE publication, one signed by all its members, suggested that more than 80% of the human genome is functional (The ENCODE Project Consortium, 2012) based upon evidences from multiple experimental protocols including RNA sequencing, binding by DNA binding proteins, DNase I hypersensitivity, histone modification, DNA methylation, and chromosome conformation capture (Cheng et al., 2012; Howald et al., 2012; Sanyal et al., 2012; Thurman et al., 2012; Yip et al., 2012). Further, the ENCODE results provided a strong evidence, accrediting vital regulatory roles to the lncRNAs predicting them to encompass much of the genome length (Derrien et al., 2012; Djebali et al., 2012). Parallel to the ENCODE the FANTOM project was initiated with the aim to

generate the complete map of the mouse transcriptome, based upon an extensive collection and annotation of full-length cDNAs. The FANTOM1 project generated 21,076 cDNA sequences, the single largest dataset of sequences coming from a given organism at its time of publication (Kawai et al., 2001). The FANTOM2 project further improved upon the experimental methods and annotation pipelines of FANTOM1 to generate 60,770 cDNA sequences of which 20% were predicted to be non-coding in nature (Okazaki et al., 2002). The publication of the FANTOM2 results coincided with the publication of the mouse genome sequence (Mouse Genome Sequencing Consortium et al., 2002) and marked an important milestone in understanding the transcriptional diversity in mammalian genomes. Further continuation of cloning and sequencing identified an additional 42,031 cDNA sequences by the FANTOM3 project (Maeda et al., 2006). Thus in total the FANTOM consortium generated ~100,000 mouse cDNA sequences of which only ~50% were annotated to be protein-coding in nature, providing a conclusive evidence for wide spread non-coding transcription. Further, utilising the CAGE technology (Kodzius et al., 2006) to map and quantify the presence of transcriptional start sites (TSS), FANTOM3 generated a map of promoter usage in the mouse transcriptome (Carninci et al., 2006). The results demonstrated the presence of tissue specific alternative TSS for a majority of protein-coding genes, allowing for the first time a genome-wide analysis of transcription initiation events in relation with tissue specificity. The Ensembl project (http://www.ensembl.org/) was launched as a database resource to store information on genes, proteins, conservation metrics and regulatory features of large genomes (Hubbard et al.,

2002). The current Ensembl version (74) supports the genomic datasets of 77 species (60 chordate, 17 non-chordate species) which include human along with commonly studied model organisms like mouse, zebrafish, *Drosophila*, *C.elegans* and yeast. The annotation and assimilation of all the data within the Ensembl databases is reliant on the Ensembl pipeline, based on Perl language modules, which retrieve data, perform analyses and submit results into the database system (Potter et al., 2004). Further, Ensembl provides its end users with computational know-how a set of software libraries written in Perl which can be used to gain programmatic access into the Ensembl databases (Stabenau et al., 2004). Apart from the software libraries Ensembl also provides a user interface to retrieve data for the general user known as the Ensembl BioMart (Kinsella et al., 2011). Recently Ensembl has also started incorporating RNAseq data in the databases, thus allowing the end user to compare the expression of various gene models across developmental stages or tissues in different organisms. An important step towards this direction is the generation of ~25,000 gene models for zebrafish from five tissues and seven developmental stages, which were used to improve the structure and annotation of the existing Ensembl gene models (Collins et al., 2012). An important component of the Ensembl pipeline is the annotation of non-coding RNA which includes small as well as long non-coding RNAs (http://www.ensembl.org/info/genome/genebuild/ncrna.html). The current version of Ensembl database has predicted lncRNAs from the human, mouse and zebrafish genomes. While the human lncRNA dataset present in Ensembl is representative of the human lncRNA catalog generated by GENCODE, the mouse and zebrafish datasets exist as

independent predictions.

1.7.2 Large-scale identification of lncRNAs

Apart from the ENCODE, FANTOM and the Ensembl consortiums a few other published reports have identified mammalian lncRNAs. More than a thousand lncRNAs were identified through an *ab-initio* sequence assembly pipeline in the mouse embryonic stem cells, neuronal precursor cells and lung fibroblasts (Guttman et al., 2010). The authors developed a program (Scripture) to reconstruct the transcriptome *ab-initio* with information of the mapped reads and the mouse genome sequence. An alternative algorithm which can emulate the substantial complexity of the eukaryotic transcriptome during the assembly of small read sequences is "Cufflinks" (Trapnell et al., 2010). The program identified ~3,700 previously un-annotated transcripts from mouse myoblast cell-lines and is cited in multiple RNAseq studies of organisms with a reference genome. A catalog of 8,000 lncRNAs were reported using both "Cufflinks" and "Scripture" to assemble RNAseq data across 24 tissues and cell types in human (Cabili et al., 2011). The results attracted attention as they combined the available lincRNA population with new candidates and the dataset was claimed to be the most comprehensive set of lncRNAs in humans when published. Alternative approaches like that of the Trinity software can perform a transcriptome assembly without a reference genome and can possibly lead to discovery of many novel lncRNAs skipped by reference genome based assembly methods due to lack of proper reference, presence of repeats and complex splicing patterns (Grabherr et al., 2011). Though

there are sporadic reports of characterised lncRNAs in diverse organisms, seldom is reported their large scale identification beyond the dimension of mammalian genomes. The number of lncRNAs identified in different organisms, including mammals is summarised in **Table 1.2**.

Organism	Number of lncRNAs	Reference
Human	13,870	GENCODE v19 (Derrien et al., 2012)
	8,000	(Cabili et al., 2011)
Mouse	4,078	Ensembl v74 (Flicek et al., 2013)
	2,740	(Guttman et al., 2009, 2010)
Zebrafish	1,754	Ensembl v74 (Flicek et al., 2013)
	1,133	(Pauli et al., 2011a)
	691	(Ulitsky et al., 2011)
Chicken	251	(Li et al., 2012b)
Drosophila	1,119	(Young et al., 2012)
C.elegans	170	(Nam and Bartel, 2012)

Table 1.2 Number of lncRNAs identified in different organisms

This dearth of datasets from different phyla exists as a major bottleneck in the field of lncRNAomics. Yet a few studies have given us insights into lncRNA dynamics in other vertebrates and invertebrates. Approximately a thousand lncRNAs were identified in the early zebrafish developmental stages showing a low level of sequence conservation at par with intronic regions, lower expression levels in comparison to coding genes and chromatin signatures resembling genes involved in development (Pauli et al., 2011a). Another 500 lncRNAs were identified using a combined strategy of chromatin marks, poly(A)-site mapping and RNA-Seq data in the zebrafish genome (Ulitsky et al., 2011). A minute fraction showed sequence conservation with mammalian lncRNAs while most of them showed a preference

to lie near *trans-dev* genes. Avian lncRNAs (~ 250) were reported from an RNAseq experiment with structural features similar to their mammalian and teleost fish counterparts but without any sequence conservation (Li et al., 2012b). Beyond the vertebrate genomes lncRNAs are demonstrated to be present in nematodes and insects. Around 1,100 lncRNAs were identified in the *Drosophila melanogaster* genome (Young et al., 2012) from RNAseq data of the modENCODE project (modENCODE Consortium et al., 2010). Interestingly though the fly lncRNAs are smaller in size as compared to their mammalian counterparts they seem to be better conserved at sequence level within the *Drosophila* clade with sequences evolving faster than ORFs but slower than UTRs. A smaller population (170) of lncRNAs were discovered in *C. elegans* where the authors state the limitations of cell type/tissue specific transcriptomics datasets for the small limited number of transcripts identified (Nam and Bartel, 2012). Interestingly, many lncRNAs reported in the mammalian brain (Chodroff et al., 2010; Ponjavic et al., 2009) were found to be extensions of 3'UTRs of coding genes using an alternative polyadenylation mechanism which was further reported to be a common observation in mammalian RNAseq data (Miura et al., 2013). This report encourages a rethink on the current lncRNA discovery pipelines from RNAseq data which may result in mis-annotation of alternative polyadenylated transcripts (APAs) of coding genes as lncRNAs. Hence it is with utmost concern that lncRNAs must be predicted considering that being non-coding and not overlapping a coding gene is not sufficient to annotate them as lncRNAs. However generally the lncRNAs which do not overlap the coding space of a genome are termed as long

intergenic non-coding RNAs (lincRNAs) (Cabili et al., 2011; Khalil et al., 2009; Ulitsky et al., 2011; Young et al., 2012) while those which are transcribed antisense to coding genes are called antisense transcripts (AS) (Faghihi et al., 2010; He et al., 2008; Katayama et al., 2005). There can be also intronic lncRNAs and sense overlapping lncRNAs but the lincRNAs and the AS together comprise the majority of the population. The lincRNAs can be further divided into divergent, convergent and 3'/5' proximal depending on their orientation with respect to the closest coding gene. In a recent study it was reported that the largest fraction of lncRNAs in human and mouse exists in divergent pairs with a coding gene, possibly sharing a promoter thus showing expression correlation with the coding gene during ESCs differentiation (Sigova et al., 2013). It suggests that the onset of transcription for a majority of lncRNA transcripts is coordinated with a coding gene in mammalian genomes. The difference in lncRNAome size in diverse organisms demonstrate that they evolve rapidly as against coding genes which are almost all conserved amongst vertebrates and bear ancestral gene associations with invertebrates. Yet it must be noted that even after undergoing a rapid evolution lncRNAs possess a faint but detectable signature for natural selection. Hence they show a low sequence conservation between closely related species and a complete absence of homology between candidates far away in the evolutionary ladder.

1.8 LncRNAs in the post-ENCODE era

It is important to focus on the status of the lncRNAs in the post-ENCODE era, keeping in mind the current resources available and those which are lacking. A

recent review curtly specifies that the ENCODE project has written a eulogy for the concept of *"junk DNA"* (Pennisi, 2012). Yet a major question which remains unanswered is whether we possess the most comprehensive catalog of human lncRNAs. Recent reports based on statistical estimation (Managadze et al., 2013) and combinatorial analyses involving multiple RNAseq datasets (Hangauer et al., 2013) suggest otherwise, predicting that only 1/5th (~ 10,000) of the total possible lncRNA genes are currently reported by ENCODE. Whether or not all of these genes confer a functionality important during cellular development and differentiation is still under debate. This simply means whether lncRNAs comprise a much larger part of the genome than currently expected, which is important for the organism survival and reproduction. We are currently way behind in experimental validation of lncRNAs in comparison to the rate of their identification in different organisms. Computational methods to measure lncRNA conservation may provide a valuable insight on their functionality. The hurdles on this path is that a few lncRNA sequences under a purifying selection are conserved on an evolutionary timescale. A fact highlighted in a report showing conserved regions in mammals with increasing diversity within humans (likely to be nonfunctional) and mammalian non-conserved regions with reduced within-human diversity procuring a novel function (Ward and Kellis, 2012). A recent review comments that the ENCODE results must help differentiate between the "identification of functional elements per se from the ascription of specific functional activities" (Stamatoyannopoulos, 2012), which elementarily suggests to avoid associating functionality to a molecule type by observing causative effects of

its sub population. The principle of gene regulation is commonly associated with lncRNAs in general, assumed to be the function of the majority of RNA molecules under this class (Barry and Mattick, 2012; Rinn and Chang, 2012). Even though there is a wide spread evidence supporting this argument the following statement effectively cautions us against extrapolation *"Moreover, the word 'regulation' has itself degraded through use by genomicists, from designating evolved effects shown or likely to enhance fitness, presumably by efficient control of the use of resources, to more broadly denoting any measurable impact of one element or process on other elements or processes, regardless of fitness consequences"* (Doolittle, 2013). Thus, it is imperative to assign functions to lncRNAs based on indirect evidences of conservation like synteny, expression correlation and mRNA secondary structure. It is not expected that such associations will cover the whole lncRNA population considering they are noted to be evolved in a lineage specific fashion, yet these approaches can fill a void which exists in understanding their evolution and plasticity.

1.9 Aims and strategies of my PhD

The aim of my PhD is to gain insights into the evolution and the functions of lncRNAs computationally and the usage of large scale functional genomics data. During these years of study and work I have developed a holistic understanding of the various computational aspects involved in the identification of long non-coding RNAs (lncRNAs) and their conservation between different species of interest. This includes the development of novel pipelines and protocols for

identification of lncRNAs and prediction of their sequence and positional conservation in multiple species. Further I have used the computational tools and strategies developed on novel RNAseq datasets to estimate the putative lncRNA populations. I have developed a computational pipeline (Annocript) which is able to annotate the coding and non-coding genes in given sequence dataset and finally predict the potential lncRNAs. The development of the pipeline is the first approach of its kind and remains currently the only software program with the capability of predict both coding and lncRNAs along with other classes of non-coding RNAs in any given dataset. I have used Annocript to demonstrate an over-estimation of the number of predicted lncRNA sequences in prior publications. Further, I have also used the Annocript to annotate the *de novo* transcriptomes of an echinoderm, a mollusc and two diatom species. I have addressed the issue of sequence conservation in long non-coding RNAs through a computational protocol which predicted a small percentage of mouse lncRNAs to show conservation in the zebrafish genome. The general lack of sequence conservation in lncRNAs over large evolutionary distances such as that between mammals and fishes, led me to search for conservation of microsynteny in lncRNAs. For this, I developed the SynLinc pipeline to identify putative microsyntenic lincRNAs between any two species with a sequenced genome and annotated transcriptomes. Comparing previously published lncRNA datasets in human, mouse and zebrafish, SynLinc was able to identify a few hundred lincRNAs, which remained closely linked across evolution with their flanking coding gene. Such an association implicates a possible co-regulation of these lncRNAs and their coding

genes. Further I have identified and analysed the specific expression patterns of lncRNAs in two novel RNAseq datasets, representing samples from a specific cell type (islet cells in zebrafish) and specific developmental stages (early development in the spotted green pufferfish). I have relied extensively on the Annocript pipeline to predict lncRNAs in both the RNAseq datasets. I defined a specific strategy for the mapping and assembly of RNAseq data which accounts for sequencing read ambiguity and is effective for the downstream identification of lncRNAs with high sensitivity. I employed this strategy to assemble the zebrafish pancreatic islet cell transcriptome followed by identification of coding and lncRNA genes. A few of the candidate islet specific lncRNAs are currently being validated in the laboratory of my external supervisor. Finally I have analysed the early developmental transcriptome of the spotted green pufferfish (*Tetraodon nigroviridis*) to show the transcriptional dynamics of both the coding and the long non-coding transcripts during early development. Further I have used the SynLinc pipeline to define a specific subset of developmentally expressed lincRNAs which remain positionally conserved with predicted lincRNAs in vertebrates. The work presented in my thesis can be divided into i) the development of computational tools for lncRNA prediction and classification ii) evolutionary conservation of lncRNAs iii) identification of lincRNAs in a specific tissue and developmental stages and prediction of their association with various biological processes. Thus, I have focused on three diverse avenues in the field of lncRNAomics and to the best of my knowledge the combination of software pipelines, conservation metrics and predicted lncRNA datasets presented here remain unprecedented.

Chapter 2

Annocript: A computational framework for annotation of transcriptome datasets and prediction of long non-coding RNAs

2.1 Introduction

2.1.1 Annotation of nucleotide and protein sequences

A primary task in the field of molecular biology is to assign a name to a gene on the basis of its function. The task of naming and assigning a functional property is more commonly termed as gene annotation. In the past annotation of a gene or a protein was dependent on repeated experiments which proved to be a time consuming process (Dearry et al., 1990; Mayo et al., 1985; Nakamura et al., 1989). The information on the annotated genes were gradually amassed in nucleic acid and protein sequence databases like the GenBank (Burks et al., 1985) and UniProt (Apweiler et al., 2004). The development of the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) provided a quick and reliable approach to annotate novel nucleotide and protein sequences based on their sequence homology with known genes/proteins in the sequence databases. A surge in high-throughput genomic technologies (Camargo et al., 2001; Okazaki et al., 2002) resulted in a data explosion of sequence information which requires sophisticated computational

pipelines for management and annotation. The Ensembl, UCSC and the FANTOM projects were the first to automate the annotation of a large number of sequences using a relational database management system and computational annotation pipelines (Hubbard et al., 2002; Kawai et al., 2001; Kent et al., 2002). The annotation pipeline were designed to gather information from multiple experiment platforms, write the primary information content along with derived meta-information to a database, which is then made available to the end-user in an organised format. The task of annotation was further complicated by the onset of next-generation high-throughput RNAseq technology (Mortazavi et al., 2008). Reduction in sequencing costs and improved detection of transcription surpassed past estimates of data generation and size of transcriptomes (Schuster, 2008). The majority of the data generated by RNAseq belong to model organisms which have the advantage of a sequenced genome and well annotated gene models (Graveley et al., 2010; Harrow et al., 2012; Harvey et al., 2013). However transcriptomes from non-model organisms are being reported in increasing numbers (Ji et al., 2012; Liu et al., 2013; Rokyta et al., 2012; Sadamoto et al., 2012; Zeng et al., 2013) which come with little or no a priori sequence information. The steep incline in generation of sequences followed by development of various annotation softwares has turned the annotation of a nucleotide/protein sequence into a multi modal task comprising the identification of sequence homology, domain signatures and functional term associations. The annotation process relies profoundly on comparative analysis with public databases like Genbank and UniprotKB (Apweiler et al., 2004; Burks et al., 1985) followed by a collation of the annotated sequences with Gene ontology

(GO) terms and enzyme classes (KEGG) (Ashburner et al., 2000; Kanehisa and Goto, 2000). In case of a large scale study consisting of thousands of transcripts, organizing the sequences, annotations and functional associations for each analysis becomes a complex task. Such tasks are more suitable to be performed by automated pipelines, hence reducing the time constraint and probability of human induced error.

2.1.2 Automated pipelines for annotation of large sequence datasets

Several software solutions have been developed by various groups (Gotz et al., 2008; Koski et al., 2005; Philipp et al., 2012; Schmid and Blaxter, 2008) to cater for this problem (annot8r, Blast2GO, autoFACT, T-ACE). These pipelines take a list of sequences as input and provide the end user with a tabular output of the annotations. The basic premise behind the annotation is a BLAST comparison against nucleotide and protein databases followed by search of conserved domain profiles and association with GO and KEGG terms. Various computational pipelines follow different methodologies to achieve the objective of annotating each query sequence. The Blast2GO and T-ACE softwares provide a user interface along with a remote database connection. The autoFACT and annota8r pipelines require the sequence databases and homology search programs to be downloaded and configured locally and provide results in text and HTML format. Though this task is performed only once, it may pose difficulties for a user/scientist lacking computational knowhow. The T-ACE software goes a step ahead of the other pipelines by integrating the RNAseq expression information of the assembled

transcripts with their annotation. The above-mentioned annotation pipelines can effortlessly compile results from multiple analyses but suffer from specific drawbacks. The Blast2GO and T-ACE pipelines rely on remote connections to public databases for homology searches to avert the download and configuration of databases and homology search programs. This is a time consuming process for current *de novo* sequencing projects exceeding 10,000 sequences and relies heavily on an uninterrupted network connection. However the remote connections can be avoided for both the programs if prior formatted BLAST results are provided locally which may not be a trivial task for a user without a computational background. The annota8r software solves this problem by installing and comparing against a local database. It relies only on UniprotKB to make such comparisons which cannot annotate sequences with distant homologies. This drawback is partially overcome by autoFACT, which allows a local installation of both protein and domain profile databases for homology search. The default run of the pipeline compares all query sequences against three protein sequence databases (Uniref90, NCBI NR, NCBI COG) followed by comparison against protein domain profiles (Smart, Pfam). The multiple comparisons increases the run time of the pipeline substantially without possibly adding to the quality of the annotations. Finally all the above-mentioned pipelines are focused towards annotating the coding part of the transcriptome and hence are unable to predict putative long non-coding RNAs.

2.1.3 Computational annotation of long non-coding RNAs

In the past few years long non-coding RNAs have been identified from RNAseq studies in a diverse group of model organisms (Cabili et al., 2011; Nam and Bartel, 2012; Pauli et al., 2011a; Young et al., 2012). Recently they were even identified in the *de novo* transcriptome of an intracellular parasite (Hassan et al., 2012). The interest of the scientific community in this class of RNAs along with the rise in number of RNAseq studies (both reference based and *de novo*) has led to the development of multiple computational solutions for their identification. The codon substitution frequency score (Lin et al., 2007) and the PhyloCSF score (Lin et al., 2011c) are two measures used by different groups to predict lncRNAs (Li et al., 2012b; Pauli et al., 2011a). Both the methods are comparative in nature and rely upon multiple genome alignments of known coding and non-coding regions to estimate a statistical phylogenetic model. Based upon the statistical model which best explains the alignment in a given genomic locus, specific sequences are then predicted as coding or non-coding. The quality of alignment is an important factor for such approaches (Schloss, 2010) which makes them more suited to perform a phylogenetic analysis on well conserved coding genes in comparison to detecting lineage specific lncRNAs with poor genomic alignments. Secondly the high computational time discourages such approaches from being integrated in an annotation pipeline. Finally alignment-based classifiers require a sequenced genome which makes them out of bounds for *de novo* generated transcriptomes lacking a sequenced genome. In contrast to the alignment based approaches, Support Vector Machine (SVM) classifiers are suited for large datasets because of

their faster computation time and partial reliance on conservation in complete genomes. Coding or Non-Coding (CONC) a SVM classifier was used to identify coding and non-coding transcripts (Liu et al., 2006) using various features like amino acid composition, peptide length, sequence entropy, secondary structure and homology to known coding sequences. This method was followed by another SVM classifier called the Coding Potential Calculator (CPC) (Kong et al., 2007) which avails the length, ORF coverage, ORF integrity and sequence homology to a protein as classifiers. These programs are trained to construct a multidimensional feature space with the classifiers on known coding and non-coding data which defines a margin between the two classes. Once trained the classifier can be used on novel uncharacterised data. The CPC and CONC softwares also suffer from long computation time for large datasets and may predict coding sequences as non-coding if they lack a known homolog in the protein databases (Kong et al., 2007). Recently published SVM based softwares iSee-RNA and Coding Potential Assessment Tool (CPAT) claim to circumvent the problem of computational time and homology accuracy by reducing the reliance on any alignment based parameter (Sun et al., 2013a; Wang et al., 2013). They employ a binary classifier between known lincRNAs (positive set) and coding transcripts (negative set) to establish a training dataset. The features used to classify are ORF length, nucleotide composition and codon usage bias. Both programs require genome wide phastCons conservation scores along with the other classifiers to build their training datasets. The phastCons program assigns a conservation score to each nucleotide base of a genome based on its alignment with other genomes (Pollard et

al., 2010). This requirement puts a limitation to their usage within non-model species without a well annotated genome. In contrast to the above mentioned programs the Portrait software (Arrial et al., 2009) requires a sequence file as the only input to report a probability score for a transcript to be non-coding. The software uses di/tri-nucleotide frequency, nucleotide sequence entropy, translated amino acid hydropathy, isoelectric point and finally length of predicted ORF as SVM classifying parameters. A major advantage of the Portrait software is that it uses the ANGLE package (Shimizu et al., 2006) to estimate the putative ORF size which is optimised to predict small ORFs. The Portrait software running time for a dataset of 4000 sequences is faster than CPC (500X) and slower than CPAT (50X) with a sensitivity at par with CPAT but lower than CPC and specificity higher than CPC but lower than CPAT (Wang et al., 2013). Five factors result in Portrait being the ideal choice for classifying non-coding transcripts in the current breed of non-coding classifiers.

- It does not need a large computation time.
- It does not require whole genome alignments to build training models like CPAT and iSee-RNA.
- It can be applied on *de novo* and reference based transcriptomes.
- It is easily integrable in an annotation pipeline.
- It has a balanced specificity and sensitivity of predictions.

The current interest in lncRNAomics has led to a community agreement of the minimum features required to annotate a sequence as a long non-coding RNA.

These are

- Sequence length greater than 200 nucleotides.
- ORF size of less than 100 amino acids.
- Lack of sequence homology with protein and domain databases.
- Lack of sequence homology with known small ncRNA classes.
- High non-coding potential (based on various properties like nucleotide composition, entropy, codon usage, ORF size, hydropathy).

2.1.4 A pipeline to annotate coding and non-coding sequences: Annocript

The currently available annotation pipelines rely on the BLAST software package to make homology searches with nucleotide, protein and domain signature databases. The homology search is the most time consuming step in an annotation pipeline. Yet none of the existing softwares are focused towards establishing a balance towards speed and accuracy of the homology search. Further, download and local installation of the databases to be used in the annotation is a complex task for an user without computational experience. An important drawback is the inability of a single automated pipeline to predict both coding and non-coding sequences in a given dataset. In light of these issues I wanted to develop a software pipeline which can accurately annotate large sequence datasets in a short period of time. Further I also wanted the pipeline to be able to predict non-coding RNAs, specially long non-coding RNAs in a given dataset. The final aim is to develop the pipeline into a community resource which can be easily downloaded and installed in any UNIX/LINUX based computer system. I developed the Annocript pipeline

for annotation of coding and non-coding sequences harnessing the combined capacity of parallel processing and multiple software packages for sequence annotation and classification. I have used Annocript extensively to predict coding and lncRNA transcripts in various projects, which are part of my PhD. I want to thank Francesco Musacchia (post-doctoral fellow in my laboratory) who took up the Annocript project post-version 1.0 and is credited for development of the pipeline to its current state.

2.2 Material and methods

2.2.1. General structure of the Annocript pipeline

All components of Annocript are implemented as Perl scripts using BioPerl modules (Stajich et al., 2002) and the Perl 5 language. Annocript runs in an UNIX/LINUX environment and requires requires a MySQL (> v5.1) account to build a database to store the source files and annotations. It needs prior installation of the NCBI BLAST+ (> v2.25) (Camacho et al., 2009), Portrait v1.1 (Arrial et al., 2009) and the Virtual Ribosome v1.1 (Wernersson, 2006) softwares. The sequence databases and mapping tables required by Annocript are downloaded directly by the pipeline. The download requires a few hours but needs to be performed only once. The files obtained from various resources are:

- protein sequences in FASTA format for the Uniref90 and SwissProt databases along with functional associations of protein sequence identifiers

with gene ontology terms (UniProt Consortium, 2009).

- protein domain profiles from the NCBI Conserved Domain Database (Marchler-Bauer et al., 2012) formatted for the rpstblastn program.
- Associations of protein sequence identifiers with enzyme identifiers from the Expasy proteomics resource (Artimo et al., 2012).

The pipeline compares query sequences against known proteins and domains, associates GO terms and enzyme IDs based on the classification of the best blast hits and calculates the probability of a sequence to be long-non coding followed by generation of text and graphical output (**Figure 2.1**). All the functions are performed by four programming modules which are involved in: 1) database creation (DB_CREATION), 2) program execution (PROGRAM_EXEC), 3) parsing of results (GFF_AND_OUTPUT) and 4) generation of statistics (OUTPUT_AND_STATS). The first module (DB_CREATION) downloads all the sequence databases and mapping tables required to perform the annotation tasks. This module needs to be executed only once since the files downloaded will be parsed and used to populate a specific database that can be used in subsequent analyses until there will be need for an update.

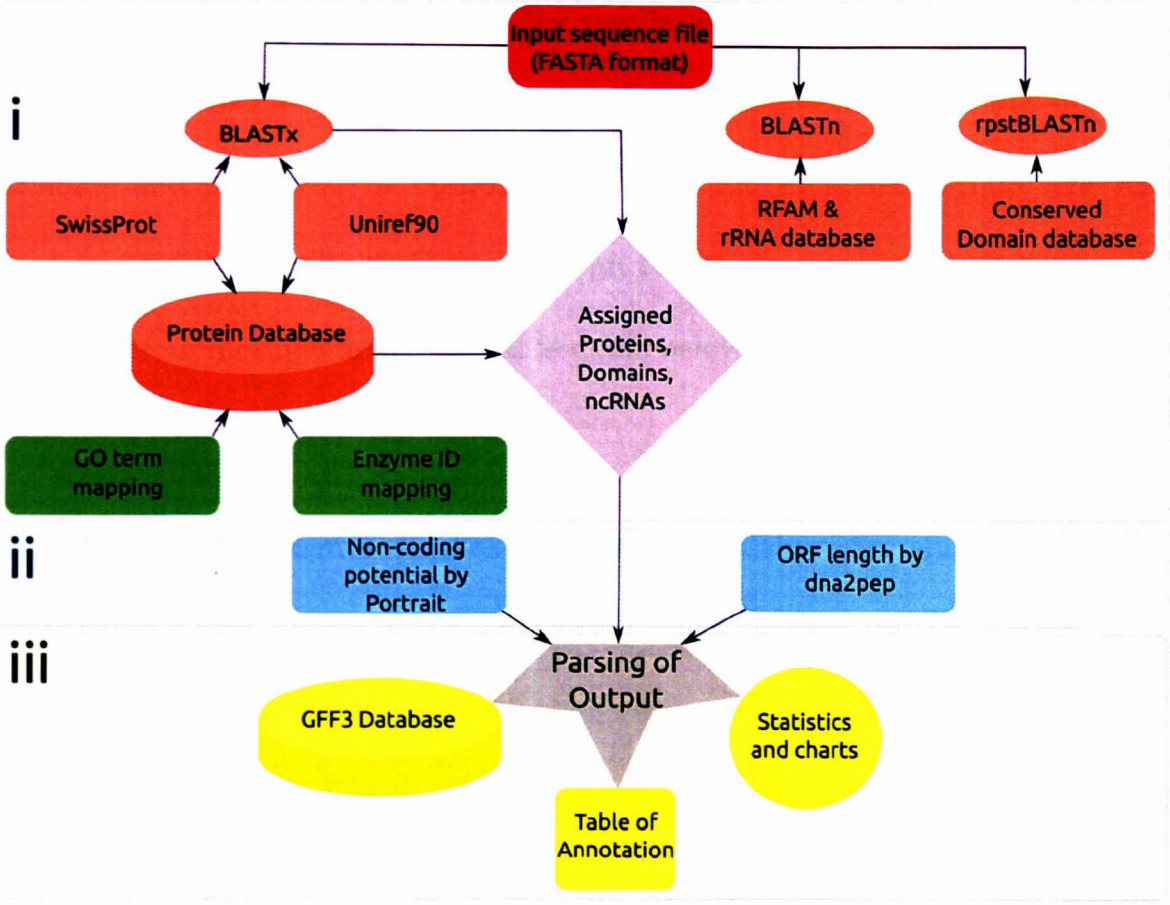


Figure 2.1 Schematic overview of the Annocript pipeline. **i)** The homology section creates a protein database from Uniref90 and SwissProt along with the association of each protein to GO terms and enzyme IDs. The database is used to quickly retrieve annotation information for each protein putatively assigned to a query sequence. The query sequences are compared against Uniref90, SwissProt, NCBI CDD and Rfam + rRNAs using the BLAST software package. **ii)** The sequence feature section obtains the longest ORF and non-coding potential score for each query sequence. **iii)** The results from the homology and sequence feature sections are parsed into GFF3 format and uploaded in a GFF database for quick multiple retrievals. **iv)** All the results are combined in a single tab delimited text file along with a HTML format output containing the statistics of annotation.

The PROGRAM_EXEC module performs a homology search of each query sequence against the downloaded databases followed by calculation of the longest

ORF size and non-coding potential. The results of the PROGRAM_EXEC module are passed to the GFF_AND_OUTPUT module which inserts the results into the database and generates the output files in text and tabulated format. The last module (OUTPUT_AND_STATS) creates an HTML document with statistics and plots. The four modules are discussed in detail below.

2.2.2 Parsing of sequence database headers and their conversion into BLAST compatible binary format (DB_CREATION)

This module downloads the following databases:

- Uniref90: ftp://ftp.uniprot.org/pub/databases/uniprot/ current_release /uniref/uniref90/uniref90.fasta.gz
- SwissProt: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz
- CDD: ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian/ Cdd_LE.tar.gz
- RFAM: ftp://ftp.sanger.ac.uk/pub/databases/Rfam/CURRENT/ Rfam.fasta.gz
- Uniprot ID mapping: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz
- Enzyme ID mapping: ftp://ftp.expasy.org/databases/enzyme/ enzyme.dat

The FASTA file headers of the Uniref90 and SwissProt databases are parsed to extract information about the protein name, description and species. Each protein is given a unique ID and is associated with GO terms and Enzyme IDs from the downloaded mapping files. All the information is uploaded into a MySQL database with the protein IDs as primary reference. The database creation reduces the time spent on parsing of data after each annotation as information can be

retrieved repeatedly without a time constraint. The sequence files downloaded in FASTA format are converted to BLAST compatible databases using the makeblastdb program of the BLAST+ suite.

2.2.3 Execution of the Annocript programs (PROGRAM_EXEC)

Although Annocript is highly configurable by the user and the majority of parameters can be adjusted in a specific configuration file, here I give the parameters currently used in my analyses. These parameters are tuned to perform an analysis with the best sensitivity and speed. The pipeline runs two principal sections of annotations based on homology and sequence features. In the homology section putative gene names are assigned to all query sequences by using the BLASTx algorithm (parameters: word_size = 4 evalue = 10^{-5} num_descriptions = 5 num_alignments = 5 threshold = 18) against the Swiss-Prot and UniRef90 databases. Protein domains are identified by running a rpstBLASTn search (parameters: evalue = 10^{-5} num_descriptions = 20 num_alignments = 20) against CDD profiles and finally a search against other non-coding RNAs (rRNA, tRNA, snRNA, snoRNA, miRNA) is done by performing a BLASTn search (parameters: evalue = 10^{-5} num_descriptions = 1 num_alignments = 1) against an integrated database of RFAM and NCBI Refseq ribosomal RNAs. The sequence feature section includes running the Virtual Ribosome program (dna2pep script) to identify the longest ORF by searching across all reading frames without explicitly looking for a start codon (paramaters: -o none -r all). The final step of this section is the calculation of non-coding potential for all the input sequence data with the

Portrait software. All sequences larger than 200 nucleotides without any homology based annotation containing an ORF smaller than 100 amino acids and a non-coding score greater than 0.95 are predicted as lncRNAs.

2.2.4 Parsing of the results into GFF3 and tabular format (GFF_OUTPUT; OUTPUT_STATS)

A GFF3 database is built by Annocript from the raw results of each analyis in the homology and the sequence feature sections. The results are converted into GFF3 format and uploaded in the MySQL database using BioPerl modules. In parallel a Perl hash structure is built for all the results which is used to quickly extract the complete annotation of each sequence into a single tab delimited text file along with a HTML document which gives overall statistics of the annotation. The Annocript results are divided into three sections a) GFF, b) tabular, c) graphical. The GFF output includes the results of different annotations performed by Annocript in GFF3 format (http://gmod.org/wiki/GFF3). It is a widely used file format system which can be parsed as well as uploaded in a database to make quick statistics. The tabular output is the main output of the pipeline where a row is assigned to each query sequence (**Table 2.1**). Each row has information on the assigned proteins, domains, ncRNA classes, length, longest ORF size and non-coding potential of the sequence. This file is recommended for filtering transcripts based upon different parameters. An abridged summary of the annotation is given as a HTML formatted document (**Annexure 1**). This gives information on the number of sequences annotated as coding and long non-coding, mean length and GC content. It also provides a chart of the frequency distribution of organisms to

which all assigned proteins belong. For *de novo* sequencing projects this informs the taxonomical proximity of the query sequences with protein sequences available in public databases. Further statistics on GO term biological process, molecular function, cellular class and protein domain abundances are also given in graphical format.

Name	Description
TranscriptName	The name of the transcript as given in the sequence file
TransLength	Length of the transcript
HSPNameSP	Highest scoring pair (HSP) with the least e.value given by BLASTx comparison against SwissProt.
HSPLengthSP	Length of HSP alignment
HSPScoreSP	E.value assigned to the HSP
HITLengthSP	Length of the HIT (SwissProt entry) with lowest e.value HSP
QCoverageSP	The fraction of query sequence covered by the HSP
HCoverageSP	The fraction of HIT sequence covered by the HSP
DescriptionSP	Description of the HIT
EnzymeIds	Enzyme ID corresponding to the HSP
EnzymeDescs	Descriptions of the Enzyme ID
HSPNameUf	Highest scoring pair (HSP) with the least e.value given by BLASTx comparison against UniRef90
HSPLengthUf	Length of HSP alignment
HSPScoreUf	E.value assigned to the HSP
HITLengthUf	Length of the HIT (Uniref90 entry) with lowest e.value HSP
QCoverageUf	The fraction of query sequence covered by the HSP
HCoverageUf	The fraction of HIT sequence covered by the HSP
DescriptionUf	Description of the HIT
Taxonomy	The organism from which the HIT originates
BPId	GO biological processes ID mapped to the Uniref90 HIT
BPDesc	Description of the GO biological processes ID
MFId	GO molecular function ID mapped to the Uniref90 HIT
MFDesc	Description of the GO molecular function ID
CCId	GO cellular component ID mapped to the Uniref90 HIT
CCDesc	Description of the GO cellular component ID
CDName	Top 5 domains with lowest e.value given by tRpsBLASTn comparison against CDD
CDStartEnd	Coordinates of the domain alignment with respect to the query sequence

CDScore	E.value of the predicted domains
CDDesc	Description of the predicted domains
OtherNCName	Highest scoring pair (HSP) with the least e.value given by BLASTn comparison against RFAM
OtherNCScore	E.value assigned to the HSP
OtherNCDesc	Description of the HIT (RFAM entry) with lowest e.value HSP
LongOrfLength	Length of the longest ORF
LongOrfStrand	Strand of the longest ORF
LongOrfFrame	Reading frame of the longest ORF
ProbToBeNonCoding	Non-coding potential score
lncRNA4Annocript	Final Prediction of the sequence to be coding or non-coding
Sequence	The input query sequence

Table 2.1 Names and description of each column of the Annocript tabular output

2.2.5 Comparison against a reference coding dataset and benchmarking the time required for analysis

List of refseq IDs for human protein coding genes having a single ortholog in mouse, zebrafish, *Xenopus tropicalis*, *Drosophila melanogaster* and *C.elegans* were obtained from Ensembl v74 using the Bioconductor (Gentleman et al., 2004) biomaRt (Durinck et al., 2005) package. A total of 2333 Refseq IDs were downloaded, whose sequences were obtained from the NCBI GenBank database (Burks et al., 1985) using a custom Perl script. The human single ortholog set (HSO) was annotated with Annocript twice, once using the default parameters followed by another run without using the BLASTx threshold parameter and without splitting the sequence file to run multiple instances of the rpstBLASTn program. Mapping of databases identifiers were obtained from two different sources. Ensembl IDs mapped to Refseq mRNA and SwissProt accessions were downloaded using the biomaRt software package. SwissProt accessions mapped to

UniRef90 accessions were downloaded from the UniProt database (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping). A final table of Refseq, SwissProt and UniRef90 accessions was prepared by merging the two mapping files.

2.2.6 Comparison against previously published long intergenic non-coding RNA datasets

The coordinates of the human and zebrafish lncRNAs were obtained from previously published studies (Cabili et al., 2011; Pauli et al., 2011a). Only the lncRNAs falling under the conservative set were considered in case of human. The coordinates of coding genes for human and zebrafish were downloaded from Ensembl (v73) using the bioconductor (Gentleman et al., 2004) biomaRt software package (Durinck et al., 2005). The coordinates of the lncRNAs were compared against coding gene coordinates of each species respectively using the intersectBED program of the BEDTools software package (Quinlan and Hall, 2010). The lncRNAs which do not overlap a coding gene are classified as lincRNAs in both the species.

2.3 Results and Discussion

2.3.1 Structure of the Annocript prediction system and comparison against a reference dataset

The Annocript pipeline takes as input nucleotide sequences and classifies them by homology search against public coding/non-coding sequence databases and

associates GO terms and KEGG enzyme IDs. Further based upon the sequence length, longest ORF size, homology and Portrait non-coding potential the pipeline classifies putative lncRNAs in the dataset (**Figure 2.2**).

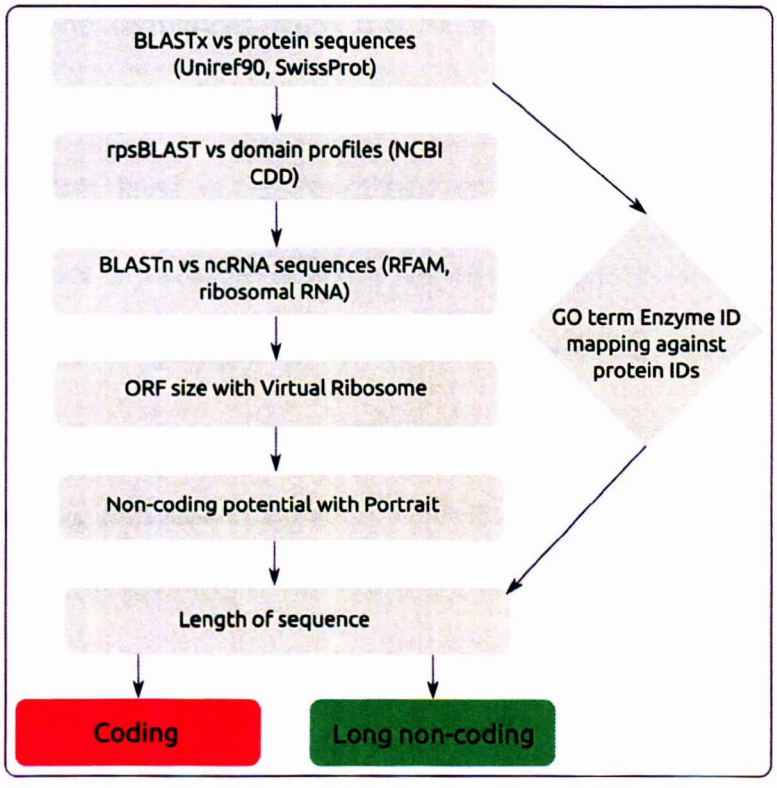


Figure 2.2 Workflow of the annotation pipeline employed to classify transcripts into coding and non-coding.

The annocript pipeline uses a combination of BLASTx search parameters (-word_size and -threshold) to significantly improve upon the runtime of the BLASTx search. This parameter is not implemented by default in other annotation pipelines discussed before. In addition, while searching for signature domain profiles, Annocript has the ability to utilise multiple processors by splitting the query sequence files and running the rpstBLASTn program independently, in parallel, on each subset of sequences. The rpstBLASTn by default cannot take advantage of parallel processing, hence this modification by Annocript enables a

faster domain search of query sequences. To test the efficiency of annotation with respect to the improvement in speed, I ran Annocript twice with and without the speed improving modifications on a reference dataset. The dataset is a set of human coding gene sequences from the NCBI GenBank refseq database (see materials and methods). These sequences are well represented in nucleotide databases with high level of conservation across multiple species. I found a significant difference in runtime between the two Annocript analyses (**Table 2.2**). After making suitable changes to increase the speed of program execution Annocript is able to run more than 10X faster. It is important to mention that currently no existing annotation pipeline is employing such modifications. Yet the improvement in speed resulted in no difference in the annotation between the two separate analyses. Further I compared the output of Annocript (default run) against accession mapping of refseq IDs with SwissProt and UniRef90 accessions. Anncoript is able to correclty identify the 83% of the sequences to their exact SwissProt accession. Another 7% of sequences are identified to their exact Uniref90 accession. The remaining 10% sequences show a mismatch between the results of Annocript and the mapped IDs coming from sequence databases. I manually inspected the Annocript results and found that all sequences are identified correctly at the level of protein names and symbols. The 10% sequences which showed a mismatch of IDs were assigned orthologs of a taxonmically close species or the protein product of an alternative isoform. Thus these results show the ability of the pipeline to maintain its sensitivity while significantly improving upon the speed of annotation.

Task performed	With modifications for speed improvement	Without modifications for speed improvement
Execution of BLASTx (vs SwissProt)	0.14 hrs	1 hr
Execution of BLASTx (vs Uniref)	2.5 hrs	24 hrs
Execution of rpstBLASTn (vs CDD)	1 hr	7 hrs
Complete execution of pipeline	3.8 hrs	33 hrs

Table 2.2 Difference in execution time of the Annocript pipeline after modifications in the BLASTx and rpstBLASTn program execution.

2.3.2 Annotation of reference lincRNA datasets from human and zebrafish

I wanted to test the ability of the Annocript pipeline to predict long non-coding RNAs. Hence I performed a default run of Annocript on previously reported long intergenic non-coding RNA sequences from human (Cabili et al., 2011) and zebrafish (Pauli et al., 2011a). At default parameters Annocript was able to predict 12 and 8% of the human and zebrafish lincRNA datasets as long non-coding (**Figure 2.3**). Such results reflect my choice to select, by default, only the most likely lincRNAs, avoiding false positives and accepting to increase the number of false negatives to come across several criticisms raised against the pervasiveness of non-coding transcription (van Bakel et al., 2010; Hüttenhofer et al., 2005). These results however suggest that there is a probable over-estimation of predicted lncRNAs in previously published datasets

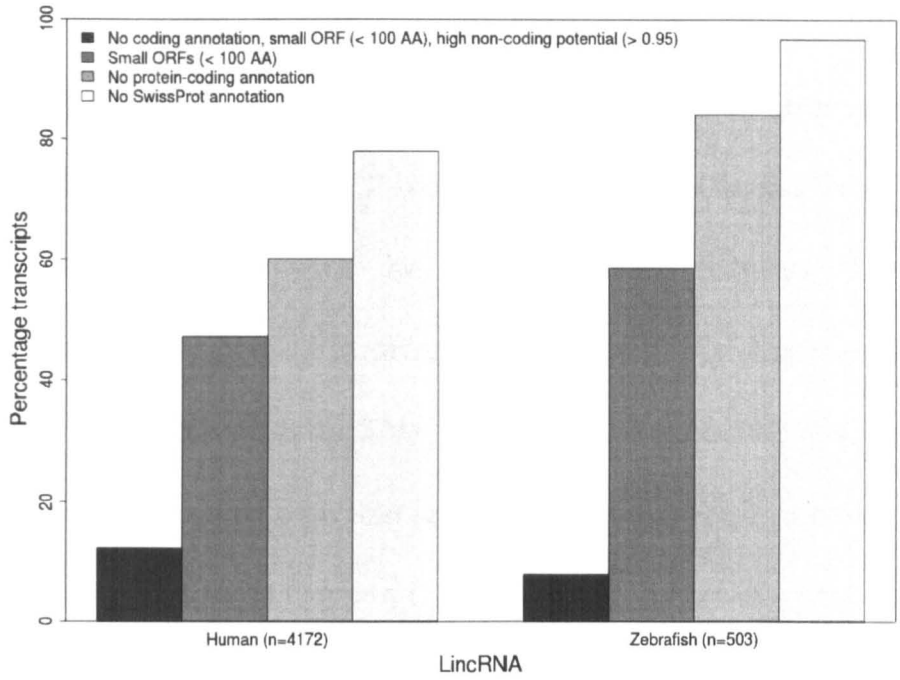


Figure 2.3 Annotation of previously published lincRNAs by the Annocript pipeline. The x axis represents the lincRNAs assigned into different categories based on Anncoript prediction. The y axis represents the percentage of each category.

2.3.2.1 Sequence and homology based strategies of Annocript

Looking specifically at the different set of evidences resulting from the Annocript analysis on the selected lincRNA datasets I can produce several considerations. While the total number of lincRNAs longer than 200 nucleotides is greater than 95%, around half of the sequences in both the datasets have a predicted ORF of less than 100 amino acids (47 and 58%). However, recent reports suggests that many non-coding transcripts are actually capable of coding for a small bio-active peptides in the mouse and drosophila genomes (Crappé et al., 2013; Ladoukakis et al., 2011). Even if a lincRNA has an ORF smaller than 100 AAs it may be classified as coding by Annocript based on Portrait non-coding potential (NCP) score threshold (15% human, 8% zebrafish lincRNAs). Apart from the ORF size the

Portrait software relies on a number of sequence composition and biochemical property based metrics to predict the non-coding potential. Thus Annocript uses both the ORF length and the NCP score together as its sequence based strategy to predict lncRNAs. Along with the sequence based features, Annocript also incorporates homology based features to predict lncRNAs. Approximately 40% of human and 16% of zebrafish lincRNA transcripts are annotated as coding on merit of their homology against a protein sequence in SwissProt, UniRef90 or a domain signature in the Conserved Domain Database. While a homology based annotation may be biased in case of lncRNAs overlapping the exons of coding genes, the datasets used here are strictly intergenic in nature. Thus the observed homology may be due to uncharacterised, predicted or hypothetical protein sequences which may also code for small peptide sequences. Indeed ~75% of human and ~50% of zebrafish SwissProt identifiers assigned to a lincRNA are classified as an uncharacterised or predicted protein having no direct experimental evidence of its translation.

2.3.2.2 Distribution of the non-coding potential scores of published lincRNA sequences

At this point I decided to compare the distribution pattern of the Portrait NCP score for lincRNAs matching a protein by the homology search. The default prediction of lncRNAs by Annocript is dependent on a highly conservative NCP score threshold of 0.95, while the authors of Portrait suggest a score of 0.5 or above to be suitable for prediction of non-coding sequences (Arrial et al., 2009). I defined

a subset of sequences based upon Annocript results, which fulfill all criteria for being a lncRNA except for the threshold NCP score (< 0.95). I address these sequences as Potential Long Non-Coding sequences (PLoNCs: Sequences > 200 nts, without homology to a coding gene with an ORF < 100 AA). I checked the distribution of NCP score for the lincRNAs predicted as coding and the PLoNCs in the human and zebrafish datasets by Annocript (**Figure 2.4**). It is interesting to note that the coding and PLoNCs subsets show a similar distribution of NCP scores for the human lincRNA dataset, while in zebrafish they show a different distribution pattern. An important point highlighted in **Figure 2.4** is the fact that sequences with homology to a predicted protein may still be given a high NCP score. It is worth noting that a subset of lincRNAs (23% human, 11% zebrafish) with an assigned protein identifier have a NCP score above that suggested by Portrait authors but below the default Annocript threshold (> 0.5 and < 0.95). This observation justifies the choice of a stringent threshold for the NCP score in Annocript. I decided to relax the NCP threshold score to the mean score of all PLoNCs predicted by Annocript in a given dataset (Human: 0.87; Zebrafish: 0.75). My aim was to reduce the stringency of predictions for datasets representing well annotated genomes such as human and zebrafish where the genomic position of a lincRNA argues against it being coding in nature. The new NCP cut-off threshold resulted in a slight increase in the number of predicted lncRNAs (Human: 25%; Zebrafish: 28%). The results indicate that the majority of the reported lincRNAs in the published studies may be potentially uncharacterised coding sequences or coding for short peptides. Unlike coding genes the computational prediction of

lncRNAs is still not well defined in terms of the principles and features which might be used for their classification.

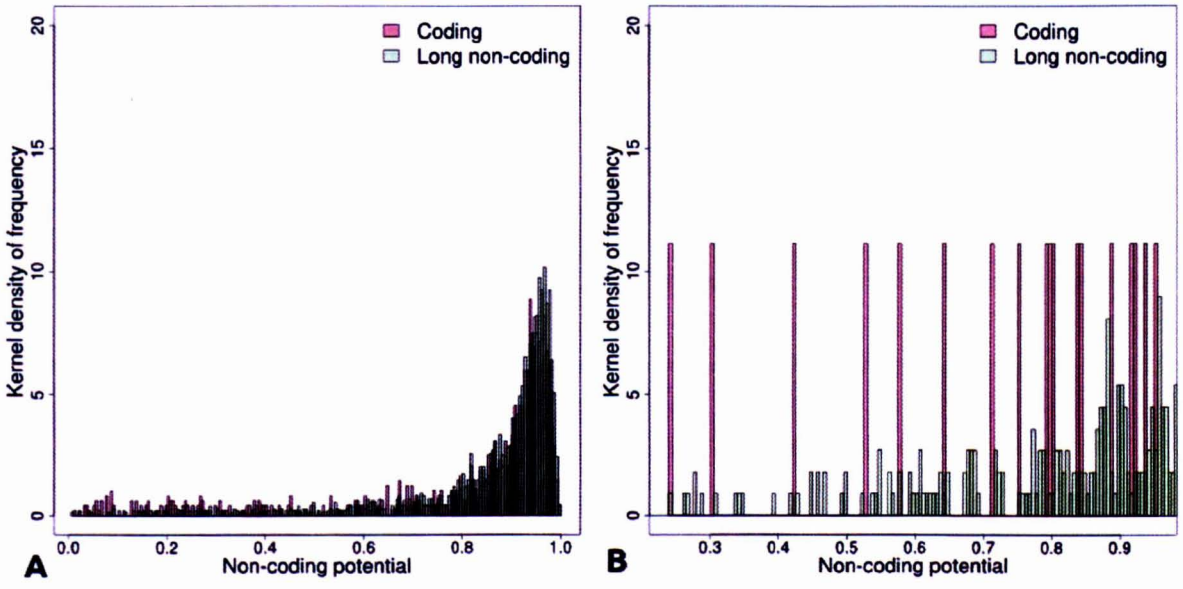


Figure 2.4 Distribution of non coding potential scores for coding and potential long non-coding sequences (PLoNCs) predicted by Annocript in **A)** human lincRNAs **B)** zebrafish lincRNAs. The x-axis represents the non-coding potential (NCP) score assigned to the lincRNAs. The y-axis represents the frequency of the the lincRNAs at a given NCP score. The bars in pink represent those lincRNAs which are predicted to be coding by Annocript, while the green bars represent the lincRNAs predicted to be Potential Long Non-Coding by Annocript.

Thus the Annocript pipeline predicts long non-coding RNAs based upon *"what they are not"* using homology (BLAST against protein databases) and sequence (sequence length, ORF size, NCP score) based strategies. Both the strategies appear to complement each other and an agreement between both ensures a high sensitivity in the prediction of lncRNAs. However the default thresholds for each filtering parameter can be altered by an end-user resulting in an increase or decrease in predicted number of lncRNA candidates.

2.3.3 Annotation of *de novo* transcriptomes using Annocript

I evaluated the ability of Annocript in handling new transcriptomic datasets. The pipeline was used to annotate the *de novo* transcriptomes of four organisms: *Astropecten aranciacus* (Starfish), *Octopus vulgaris* (Octopus), *Pseudo-nitzchia multistriata* (marine diatom) and *Seminavis robusta* (freshwater diatom). The transcriptome data used are all unpublished provided by collaborators interested in getting molecular insights in their species of interest. A special interest was in the number of potential lncRNAs predicted in each species, which currently no existing annotation pipeline is able to estimate. The source of the RNA samples are i) Whole organism early development in starfish, ii) Adult neural tissue in *Octopus* iii) Different mating types in the diatoms. The sequencing was done on the Illumina platform and assembly of reads was performed by the Trinity suite of programs (Grabherr et al., 2011). The pipeline was run with the described parameters and it was able to annotate 35-70% of transcripts in different species. Specifically 20-60% of genes were annotated against the protein coding databases while 4-12% were predicted to be noncoding (**Figure 2.5**). There is a large variation in the number of coding and lncRNA transcripts predicted in these organisms. However partially the differences can be explained by the large evolutionary distances between the chosen species. The number of non-coding RNAs is suggested to be proportional to the developmental complexity of an organism (Mattick, 2011). Hence the *Octopus* and the starfish being higher in the evolutionary ladder are expected to show more diversity in the number of non-coding transcripts in comparison to unicellular diatoms.

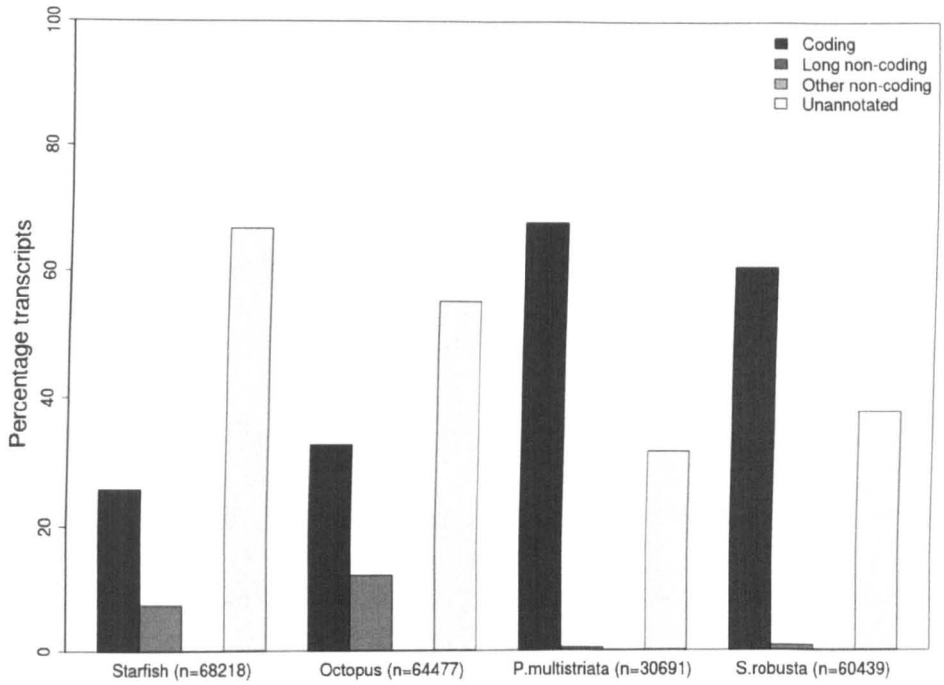


Figure 2.5 Annotation of de novo transcriptomes by the Annocript pipeline. The x-axis represents the different classes of transcripts predicted by Annocript in the assembled transcriptomes of the given species. The y-axis represents the percentage of each transcript class with respect to all transcripts for a given species.

In the *Octopus* and starfish datasets the majority of transcripts remained unannotated (55, and 67%). The annotations for *Octopus* are consistent with a recent work on the *Octopus* brain transcriptome which reports that only around 20% of the transcripts can be annotated (Zhang et al., 2012). In contrast, about 80% of transcripts were reported to be annotated in the *Strongylocentrotus purpuratus* (sea urchin) genome (Tu et al., 2012) which is the taxonomically closest species to the starfish with a sequenced genome. However a lack of sequence information in echinoderms (Kondo and Akasaka, 2012) may be responsible for the high number of unannotated sequences. The interesting aspect is the prediction of putative long non-coding RNAs in both *Octopus* (7%) and starfish (12%) since lncRNAs are not

reported in either molluscs or echinoderms till date. The percentage of unannotated genes were lower in the diatom species (*P.multistriata:* 31%, *S.robusta:* 38%) with a small fraction of genes predicted as long non-coding (*P.multistriata:* 0.42%, *S.robusta:* 0.82%). Although some of the transcripts may represent assembly artefacts we still estimate a sizable fraction to be lncRNAs. In fact the classification by Annocript enhances the repertoire of lncRNAs in non-chordates. I wanted to find the effect of NCP scores on the annotation of the given datasets. Hence I compared the NCP score distribution of all coding sequences against Potential Long Non-Coding Sequences (PLoNCs: Sequences > 200 nts, without homology to a coding gene with an ORF < 100 AA) for each of species (**Figure 2.6**). I detected that a NCP score of 0.5 is optimal to separate distribution of the coding sequences from the PLoNCs. In fact a score of > 0.5 is also recommended by the authors of the Portrait software (used by Annocript to calculate NCP score) for prediction of non-coding sequences (Arrial et al., 2009). Still it is interesting to note that the maximum density of PLoNCs are near the default Annocript cut-off of 0.95. The results from the Annocript pipeline can be filtered on a user defined choice of NCP threshold. However I have chosen a stringent default cut-off (0.95) to help the user to focus on a limited set of high scoring putative lncRNAs. This is specifically useful for the analysis of *de novo* transcriptomes since they might contain an unknown number of novel uncharacterised coding sequences. The default score ensures a minimum number of false positives in the final predicted lncRNA dataset.

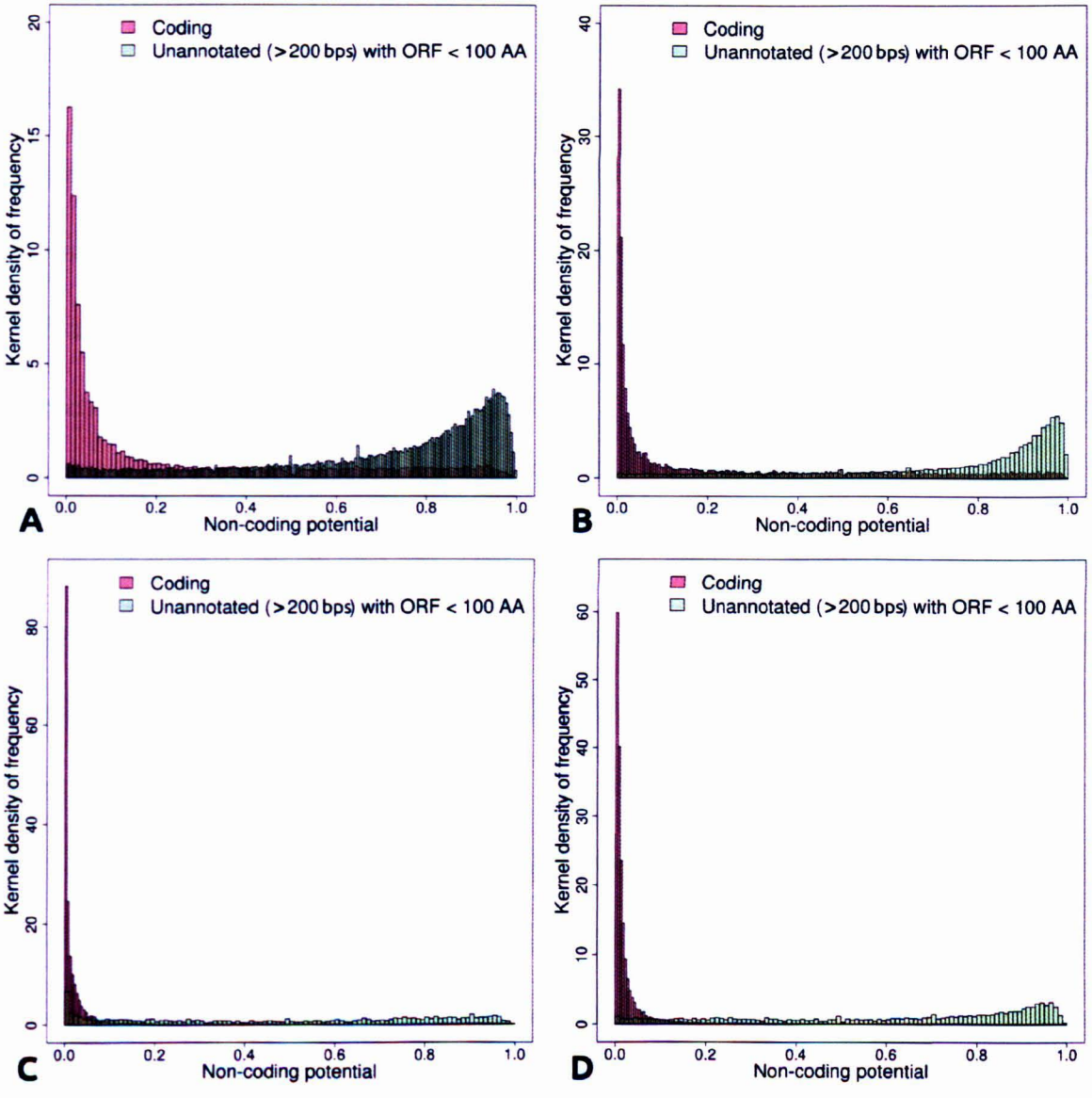


Figure 2.6 Distribution of non coding potential scores for coding and potential long non-coding sequences (PLoNCs) in **A)** Starfish **B)** *Octopus* **C)** *P. multistriata* **D)** *S. robusta*. The x-axis represents the non-coding potential (NCP) score assigned to the lincRNAs. The y-axis represents the frequency of the lincRNAs at a given NCP score. The bars in pink represent those transcripts which are predicted to be coding by Annocript, while the green bars represent the transcripts predicted to be Potential Long Non-Coding by Annocript.

2.4 Conclusion

The Annocript pipeline was developed keeping in mind two factors i) Quick and reliable annotation of large scale Next Generation Sequencing projects taking advantage of the latest softwares and computational prowess of parallel processing ii) Building a universal platform for prediction of non-coding RNAs, specially long non-coding RNAs. The pipeline takes as input nucleotide sequences and classifies them by homology search against public coding/non-coding sequence databases and associates GO terms and KEGG enzyme IDs. Further based upon the sequence length, longest ORF size, homology and Portrait non-coding potential the pipeline classifies putative lncRNAs in the dataset. The pipeline is easily configurable on a local machine and optimized to run quickly without bargaining on accuracy. The pipeline is unique in comparison to published projects of similar caliber as it can handle large datasets without requiring months in computation and is not reliant on an external server for making annotations. The prediction of lncRNAs by the pipeline is based upon sequence homology, longest ORF size and non-coding potential. It is currently the only computational pipeline capable to do so without the need of a sequenced genome or establishment of complex statistical training models. The pipeline was tested on *de novo* transcriptomes of various organisms and performed appreciably to predict of lncRNAs for the first time in these taxonomic groups. The results from Annocript classification of known lincRNA datasets reflect the fact that probably there is an overestimation of such transcripts in published studies. In this regard Annocript can be a potential benchmark in future for measuring the lncRNAome in a given trasncriptomic dataset. In

summary the Annocript is projected as a software pipeline built to provide a quick one-stop resource for annotation of the coding and non-coding sequences in large scale transcriptome projects.

Chapter 3

Sequence conservation in long non-coding RNAs over large evolutionary distances

3.1 Introduction

3.1.1 Conservation of sequence in long non-coding RNAs

Recently published studies have reported catalogs of long non-coding RNAs (lncRNAs) in a diverse range of organisms like in mammals (Aprea et al., 2013; Cabili et al., 2011; Derrien et al., 2012), a nematode (*C.elegans*) (Nam and Bartel, 2012), an insect (*Drosophila melanogaster*) (Li et al., 2009; Young et al., 2012) a fish (*Danio rerio*) (Pauli et al., 2011a; Ulitsky et al., 2011) and an amphibian (*Xenopus tropicalis*) (Paranjpe et al., 2013). A set of computational metrics are usually adopted to predict the lncRNAs which include lack of sequence homology against protein sequences, presence of small ORFs (< 100 amino acids) and a non-coding potential score. Previously I have developed a pipeline (Annocript) for prediction of lncRNAs within a dataset of nucleotide sequences. I have used Annocript to demonstrate that there is a probable over-estimation of predicted lncRNAs in previously published datasets. However, experimental validation is the foremost requirement to establish the verity of a predicted lncRNA. Yet, the evidences supporting conservation of sequence, secondary structure, location or expression

aid in identification of *bona fide* candidates from a given lncRNA dataset. The conservation of lncRNA sequence between different species is a computational aspect not explored acutely in previous studies involving prediction of lncRNAs. The primary reason is that lncRNAs showed little or no sequence conservation over long evolutionary distances. In zebrafish majority of lncRNAs are reported to have a low sequence conservation level similar to that of coding gene introns (Pauli et al., 2011a) while another study predicted a few lncRNAs (5%) to contain short stretches of sequences conserved amongst vertebrates (Ulitsky et al., 2011). An exception is the case of *Drosophila*, where more than 90% of the lncRNAs have multi species conserved elements (insects only) but there is no mention of their conservation with other vertebrates (Young et al., 2012). The principal factor associated with lack of conservation in lncRNAs is their fast rate of evolution (Pang et al., 2006) even at a small evolutionary distance. For example, variation in human lncRNA sequences are reported to comprise more than 50% of the genetic variation between human and chimpanzee genomes (Khaitovich et al., 2006). Nevertheless, in specific cases lncRNAs are reported to contain terse segments of conservation interspersed with a long span of variable sequence. A prime example is of the *Xist* lncRNA which was first reported to mediate the X chromosome inactivation in human (Clemson et al., 1996). The human *Xist* sequence is long (17 kb) but contains only few short regions (~60 bp), conserved in eutherian mammals (Duret et al., 2006b). Another example is of the *Sox2* overlapping transcript (*Sox2ot*) reported to be a key regulator of pluripotency in mouse, which contains a few highly conserved elements (HCEs) separated by large regions of low sequence

similarity (Amaral et al., 2009). Four lncRNAs described in mouse appear to be a deviation from the common pattern of low sequence conservation with more than 30% of their transcript length being conserved with orthologous sequences in chicken (Chodroff et al., 2010). The largest set of lncRNAs predicted to show sequence conservation are 659 transcripts in mouse which are constrained in their nucleotide substitution rates against the human genome in comparison to that of local ancestral repeats (Ponjavic et al., 2009). A fraction of these lncRNAs are expressed in the mouse brain (defined as *CNS-specific)* and are enriched to lie near coding genes with similar expression pattern, implicated in development and regulation of transcription. In line with the little supporting evidence, a lack of sequence conservation is generally associated with lncRNAs.

3.1.2 Protocol for identification of sequence conservation in lncRNAs

Majority of the studies have used whole genome alignments between multiple species to estimate the sequence conservation in lncRNAs (Derrien et al., 2012; Nam and Bartel, 2012; Pauli et al., 2011a; Ponjavic et al., 2007, 2009). A systematic and unbiased analysis of sequence conservation in lncRNAs is currently wanting, specially when the question is about conservation over large evolutionary distances such as between mammals and fishes. Three important factors must be considered for such an analysis

- Selection of candidate lncRNAs showing sequence conservation based upon an unbiased choice of computational parameters.
- Ability of the parameters to separate conservation in lncRNAs from

background noise.

- Definition of a specific protocol using the choice set of parameters.

I wanted to develop a pipeline for identification of sequence conservation in lncRNAs based upon the above mentioned criteria. The aim is to identify a subset of mouse lncRNAs which potentially appear to retain their function on merit of their sequence conservation with the zebrafish genome. Further I wanted to select a few exemplar mouse lncRNAs whose conserved counterparts in the zebrafish genome can be experimentally validated for functional similarity. This development of the sequence conservation pipeline and its subsequent usage to identify conserved lncRNAs in mouse, was an integral part of the work done by me in the first year of my PhD which is accepted in a peer reviewed journal (Basu et al., 2013).

3.2 Materials and Methods

3.2.1 Selection of the sequence datasets used for conservation analyses

The mouse CNS (Central Nervous System specific) and NCNS (non Central Nervous System specific) constrained lncRNA datasets were obtained from a previous study (Ponjavic et al., 2009). Ensembl lincRNA dataset was obtained from BioMart (Haider et al., 2009) and is based on the Ensembl version 62 (Flicek et al., 2011). The lncRNA sequences in each dataset were shuffled with the shuffle program (part of the SQUID C library by Sean Eddy, the executable can be found

in the HMMER3 program) (Eddy, 2011). The sequences in each dataset were shuffled 100 times resulting in three random sequence datasets rCNS, rNCNS and rEnsembl. PhastCons elements for zebrafish (zPHS) were obtained from the UCSC table browser (Dreszer et al., 2012; Pollard et al., 2010) with the "most conserved" option selected for sequence retrieval. The coordinates of the phastCons elements were mapped to the zebrafish current genome build (zv9) using the UCSC liftover tool (www.genome.ucsc.edu/cgi-bin/hgLiftOver). A total of 816,471 conserved elements were mapped out of 881,975 original elements.

3.2.2 Identification of sequence homology between lncRNAs and the phastCons elements

The lncRNAs (CNS, NCNS, Ensembl) and the random datasets (rCNS, rNCNS, rEnsembl) were compared against the zPHS using BLASTn from the BLAST+ software package (version 2.25) (Camacho et al., 2009). The BLASTn program was run with default parameters except for the word size. BLASTn comparisons with word size from 8 to 11 were performed for the CNS specific lncRNA and rCNS datasets against the phastCons elements. The NCNS/rNCNS and Ensembl/rEnsembl datasets were compared against zPHS at word size 11. I selected four BLASTn result parameters for the ROC analyses: query coverage (fraction of a lncRNA which is aligned to a phastCons element), alignment length (the length of the alignment including the gaps inserted), percentage identity (number of identical base matches between the query and the subject sequences) and e-value (a score which defines the probability of an alignment not being

random in nature). The alignments of the lncRNAs (CNS/NCNS/Ensembl) against the zPHS were taken as the true positive dataset while those from the randomized datasets (rCNS/rNCNS/rEnsembl) were considered to be the false positive set. The ROCR package in R environment was used to build the receiver operating characteristic (ROC) curve of false positive against true positive values for each parameter (Sing et al., 2005). ROC curves for the e-value parameter in the plots show the reciprocal of the e-value (1/e-value), as plotting the e-value produced curves resulting skewed below the diagonal line. Each alignment generated from the BLASTn search of the CNS dataset against zebrafish was tested for structural conservation. SISSIz program (Gesell and Washietl, 2008) was used to randomize each alignment 100 times using a dinucleotide model (SISSIz --simulate --tstv -n 100) to generate a randomized alignment dataset to measure the structural conservation (srCNS). The alignments of the CNS and srCNS datasets were checked for RNA secondary structure conservation with the RNAz 2.0 software (default parameters) (Washietl et al., 2005). To build ROC curves I used the following parameters from the RNAz output: ratio of pairwise identity by sequence conservation index, Z score and P values (1/P value). The parameters from the original alignments were considered to be true positive while those from the randomized alignments were considered to be the false positive. ROC curves of the false positive against the true positive were plotted for each parameter.

3.2.3 Identification and enrichment analysis of genomic features

The predicted conserved mouse lncRNAs were obtained using an e-value and

query alignment length cut-off allowing less than 0.05% false positives (as defined by ROC curves). The conserved lncRNAs (named cCNS, cNCNS, cEnsembl) and their respective zPHS elements sharing sequence similarity (named zCNS, zNCNS, zEnsembl) were back mapped to the mouse and zebrafish genomes (mm9 and zv9) respectively using BLASTn with default parameters but -culling_limit=1. The mapped coordinates of the mouse lncRNAs and zebrafish conserved elements were used to retrieve overlapping genes, transcripts, exons, and the closest flanking protein coding genes in a 1 megabase window using custom perl scripts utilising the Ensembl core modules API (Flicek et al., 2010) to access the Ensembl database (v62). DAVID gene annotation tool was used for the GO term enrichment and tissue expression enrichment analyses of the protein-coding genes flanking and overlapping the conserved elements using the whole transcriptome as universe (Huang et al., 2009a). An EASE score of 0.05 (Hosack et al., 2003) was used as a cut-off for the enrichment analysis. Sequences of ultraconserved elements (Bejerano et al., 2004; Sakuraba et al., 2008) were mapped on the mouse genome using BLASTn (-task blastn -culling_limit 1) with default parameters. The coordinates of the mapped elements on the mouse genome were checked for overlap with conserved mouse lncRNAs using intersectBed program from the BEDTools package (version 2.14.2) (Quinlan and Hall, 2010) with default parameters. In all the overlap analyses performed I have considered an overlap of at least 1 bp between the conserved element and the specific feature considered as sufficient.

3.2.4 Identification of orthologs between mouse and zebrafish and mapping of ESTs in the region of conservation

Zebrafish and mouse gene orthology information was downloaded from BioMart (Haider et al., 2009) based on Ensembl version 62. We collected all the Ensembl genes mapped in intervals up to 2 Mb (1 Mb up and down-stream) around each conserved element of both the genomes. For each element we looked for genes considered evolutionary related (classified as ortholog one to one, ortholog one to many or ortholog many to many) in Ensembl compara (Vilella et al., 2009). Conserved elements were considered syntenic if showing at least one evolutionary related gene in the given interval for the species considered. The analysis was performed individually on all lncRNAs stemming from the cCNS, cNCNS and cEnsembl datasets. The EST coordinates for mouse and zebrafish were downloaded from UCSC databases on 14th September 2011.

- Mouse: http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/all_est.txt.gz
- Zebrafish: http://hgdownload.cse.ucsc.edu/goldenPath/danRer7/database/all_est.txt.gz

The region of sequence conservation in the mouse lncRNAs (cCNS/cNCNS/cEnsembl) were checked for the overlap with a reported EST on the mouse genome. The same process was repeated on the zPHS conserved fragments (zCNS/zNCNS/zEnsembl) with respect to zebrafish ESTs. The Ensembl genome browser was used to generate the images for the conserved zPHS regions (Flicek et al., 2012a) and their corresponding lncRNA in mouse.

3.2.5 Mapping of RNAseq data and read count on conserved regions

The zebrafish paired end RNAseq data from 7 developmental stages and stickleback paired end RNAseq from 9 tissues were downloaded from the European Nucleotide Archive in fastq format (Accessions: SRP012923 and SRP009426). The raw reads were mapped to the zebrafish and stickleback genome using Tophat 2.0.4 (Trapnell et al., 2009) (tophat -p -o -G) and reads overlapping the conserved regions were calculated using custom Perl scripts and the coverageBed (coverageBed -split -aBam -b) program from the BEDTools package (Quinlan and Hall, 2010) (version 2.14.2). Conserved zebrafish sequences were mapped on the stickleback genome using BLASTn (-task blastn -culling_limit 1) with default parameters and a minimum 70 percentage sequence identity. Random regions (~1,200) on the zebrafish genome were selected using the shuffleBed (shuffleBed -i -g) program from the BEDTools package. Overlap associations for the random regions were calculated in the same way as that for conserved regions.

3.3 Results and Discussion

3.3.1 Selection of the mouse lncRNA datasets

The approach of my analysis is focused upon mouse lncRNAs predicted to have constrained nucleotide substitution rates amongst mammals (Ponjavic et al., 2009). The choice of the dataset reflects the fact that such transcripts comprise a subset of lncRNAs manually curated and annotated and therefore will have a lower probability to contain unannotated coding genes. Although the analysis considers

published and very well annotated lncRNAs, a few candidates can potentially code for small peptides as reported by a few recent studies indicating a few lncRNAs to be bifunctional encoding both mRNAs and functional noncoding transcripts (Bánfai et al., 2012; Dinger et al., 2011; Ingolia et al., 2011). Hence, though biological validations remain a critical factor for proper classification of these elements, based upon the choice of dataset and stringency of the analysis I am confident of the verity of my results. Further I have a a set of mouse lncRNAs, representing transcripts from different tissues and development stages predicted by the Ensembl pipeline as my second dataset (Flicek et al., 2011). I compared the mouse lncRNAs from both datasets against the zebrafish phastCons elements (Pollard et al., 2010) to predict conserved lncRNA regions in the zebrafish genome. The phastCons elements are genomic elements predicted by the phastCons program using a hidden Markov model-based method that estimates the probability of each nucleotide to be conserved based on multiple alignments of selected species. I used the phastCons6way track to select elements conserved among fishes. These are based on scores built on multiple alignment of the zebrafish genome with *Tetraodon*, stickleback, human, mouse and *Xenopus tropicalis*. These elements represent the best selection of conserved regions in zebrafish, at first instance, among fishes, but many are also conserved among vertebrates. This choice implicitly adds more genomes to our analyses and is based on the assumption that lncRNAs conserved between mouse and zebrafish are expected to be primarily conserved among teleosts. For this pilot study, the reduction in the dataset dimension, given by such choice, limited the zebrafish genomic search space to the

phastCons sequences, rather than to the full genome, making it feasible to use several randomizations steps (shuffling of the query sequences) to specifically identify the level of conservation of lncRNAs.

3.3.2 Selection of conservation parameters to identify significantly conserved lncRNAs

I developed a pipeline to identify conserved mouse lncRNA fragments in zebrafish using sequence identity, randomization and the identification of an unbiased threshold to detect significant levels of conservation (**Figure 3.1**). I used receiver operating characteristic (ROC) like analyses to select the best measures from BLASTn which help detect conservation of lncRNAs out of the following: 1) query coverage, 2) query alignment length, 3) percentage identity and 4) e-value. ROC like analyses were performed on the results of the following BLASTn comparisons: 1) mouse lncRNA against zebrafish phastCons elements (true positive set), 2) shuffled mouse lncRNA sequences against zebrafish phastCons elements (false positive set). A threshold value was defined for each parameter which resulted in < 0.05% false discovery rate (FDR). The analysis was applied on different datasets which led to the identification of 4 to 11% of the sequences in the true positive datasets to be significantly conserved. The conserved zebrafish regions show a mean length of 160 nucleotides with an average percentage identity of about 80% with their corresponding mouse lncRNA fragments (**Figure 3.1 B, C**).

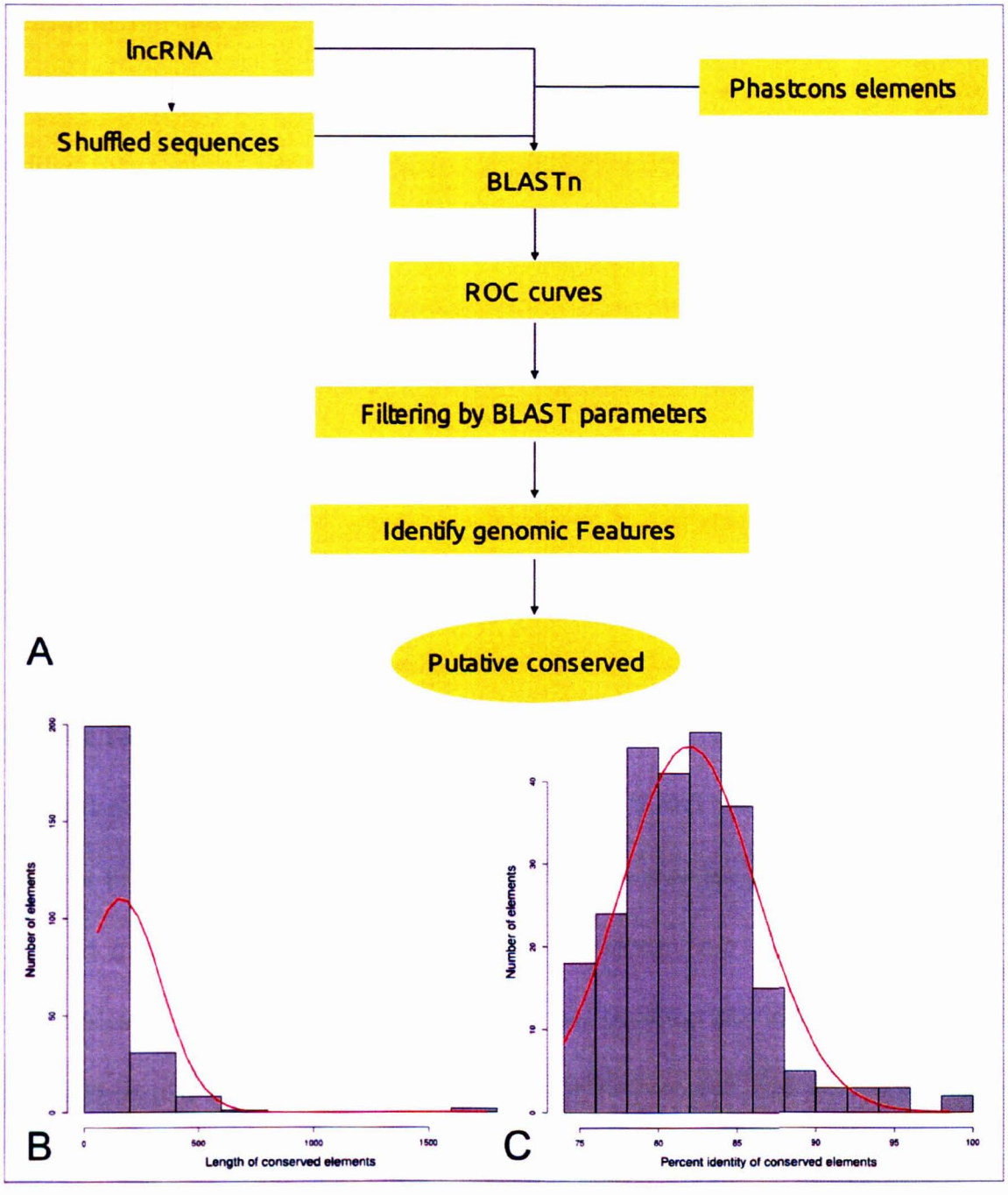


Figure 3.1 Pipeline to detect lncRNA sequence conservation and descriptive statistics of the identified conserved elements. **A)** Schematic representation of the pipeline created to identify putative conserved mouse long non-coding RNAs in the zebrafish phastCons elements. **B)** Distribution of lengths of the identified conserved elements **C)** Distribution of percentage identities of the identified conserved elements.

Mouse lncRNAs from two sources representing three datasets, were used to

determine the sequence conservation. Mammalian constrained lncRNAs from mouse (659 transcripts defined as CNS/NCNS dataset) (Ponjavic et al., 2009) were divided into Central Nervous System specific (239 CNS transcripts) and non-CNS specific (420 NCNS transcripts) along with lncRNAs identified in the mouse genome by the Ensembl lincRNA annotation pipeline (Flicek et al., 2011) (2,147 Ensembl transcripts, Ensembl version 62, http://www.Ensembl.org/info/docs/genebuild/ncrna.html). I considered the CNS lncRNAs as my primary dataset to perform an initial assessment of the BLASTn search sensitivity on the alignments between mouse lncRNAs and zebrafish conserved elements. The BLASTn word size is the parameter which calibrates its search sensitivity. The BLASTn program performs a heuristic search by locating short matches between two sequences, the length of the short match being the word size. The word size is inversely proportional to the speed of BLASTn comparisons hence a larger word size means a faster analysis. I executed multiple BLASTn runs with different word sizes ranging from 8 to 11 nucleotides on the CNS dataset. ROC curves, plotting the distributions of the indicated measures (**Figure 3.2 A**) suggest that the reciprocal of the e-value (1/e-value) is the factor capable to better segregate results between the true positive and false positive sets (area under curve, AUC = 0.79). The reciprocal is recommended when a measure produces a ROC curve significantly skewed below the diagonal line (Fawcett, 2004). In addition to the e-value I noticed, by manual inspection of results, that the alignment length (AUC 0.64) is capable of filtering low complexity (repeated) regions that may potentially align to multiple regions on the genome with a small e-value (**Figure 3.2 B**).

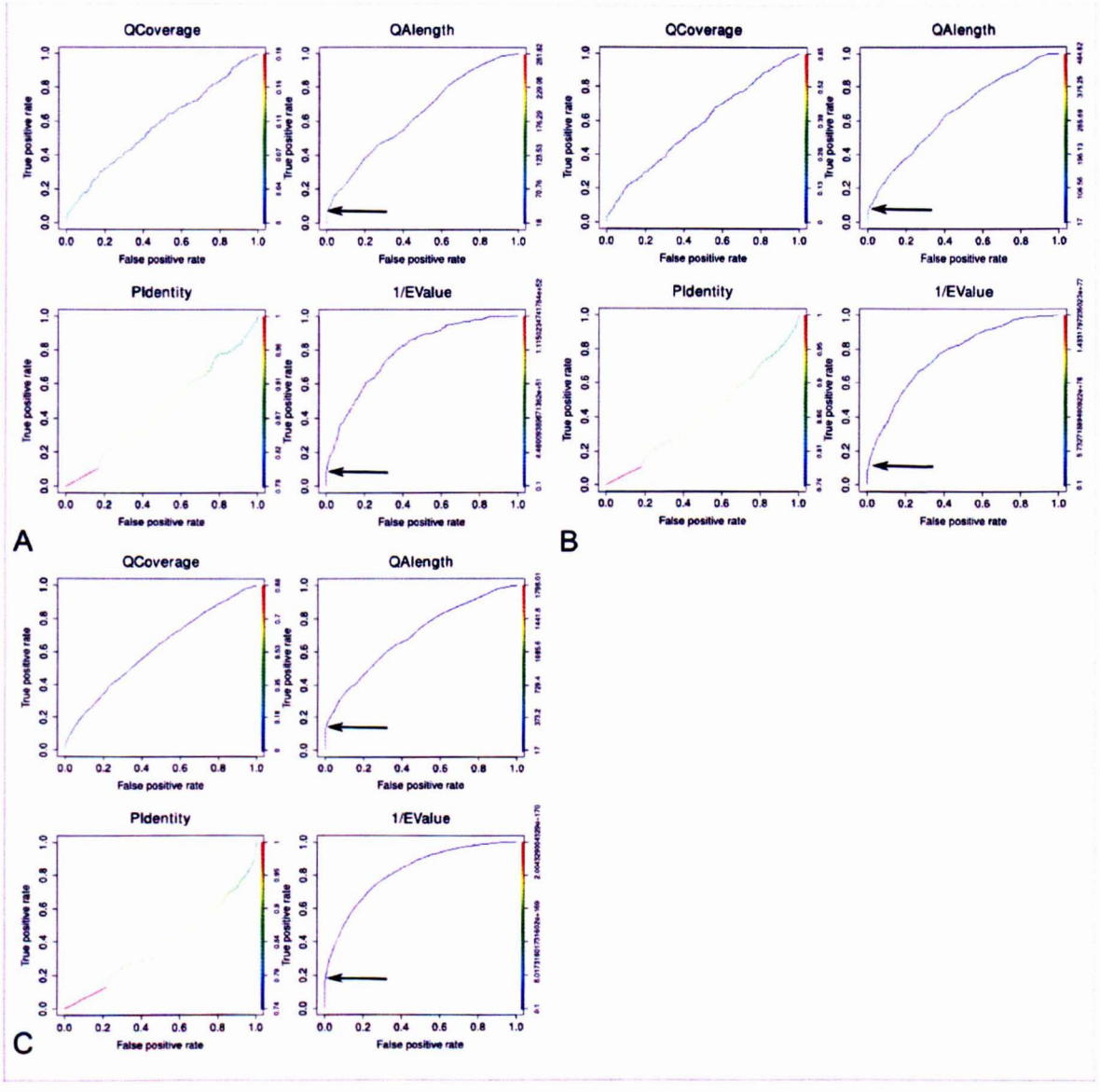


Figure 3.2 ROC curves of CNS, NCNS and Ensembl datasets homology search results. The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate for specific measures of the BLASTn results. The BLASTn search of lncRNAs against the phastCons elements represents the true positive data while the false positive data accounts for the BLASTn search of shuffled sequences against the phastCons elements. The ROC curves determine the ideal threshold which may separate the alignments with biological significance from the random occurring alignments. ROC curves for query coverage (QCoverage), percentage identity (PIdentity), query alignment length (QAlength) and e-value (1/EValue) at word size 11 for **A)** CNS dataset **B)** NCNS dataset, **C)** Ensembl dataset. The cut-off for a parameter is defined as the point of steep incline in the true positive

rate as compared to the false positive rate. The significant cut-off defined in the present analysis are indicated by arrows. ROC curves for the e-value parameter in the plots show the reciprocal of the e-value (1/e-value) because plotting the e-value produced curves sensibly skewed below the diagonal line.

It is now becoming evident that repeats are enriched in lncRNAs (Carrieri et al., 2012; Kelley and Rinn, 2012) but the presence of repetitive regions in the results reduces the specificity of predictions. Hence I decided to combine the two (e-value, alignment length) measures to select the significantly conserved lncRNAs. Indeed combining the two parameters resulted in no false positives for each dataset (FDR = 0.0%). Interestingly, the change in BLASTn word size does not affect the performance of the classifier (**Figure 3.3**). Therefore, word size of 11 nucleotides is used in all subsequent analyses.

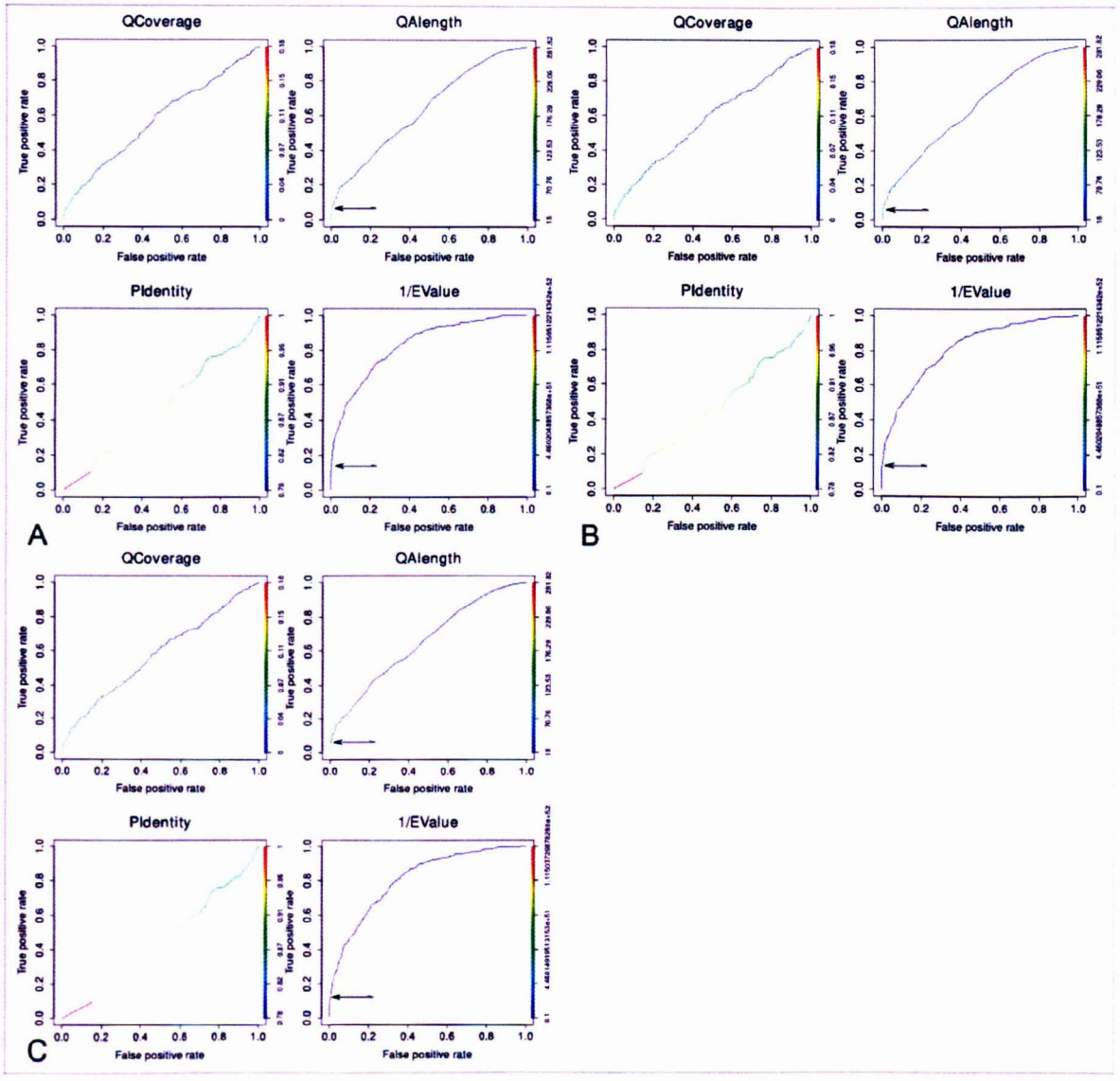


Figure 3.3 ROC curves of CNS dataset at word size 8-10. ROC curves for query coverage (QCoverage), percentage identity (PIdentity), query alignment length (QAlength) and e-value (EValue) for the CNS dataset at word size **A)** 8 **B)** 9, **C)** 10. The cut-off for a parameter is defined as the point of steep incline in the true positive rate as compared to the false positive rate. The significant cut-off defined in the present analysis are indicated by arrows.

I defined the threshold values for each alignment measure as the value at which < 0.05% false positives (randomized sequences) were predicted as conserved. An e-value cutoff of 5e-05 and an alignment length cut-off of 70 nucleotides satisfied

this criteria resulting in 11 lncRNAs from the CNS dataset significantly conserved within the zebrafish phastCons elements (**Table 3.1**).

Dataset	Word Size	Number conserved lncRNAs	% conserved lncRNA	% conserved shuffled
CNS (239)	11	11	4.60%	0.0%
	10	11	4.60%	0.0%
	9	11	4.60%	0.0%
	8	11	4.60%	0.0%
NCNS (420)	11	23	5.40 %	0.0%
Ensembl (2,147)	11	250	11.6%	0.0%

Table 3.1 The number of lncRNA putatively conserved in each dataset (CNS, NCNS, Ensembl) after applying the query alignment length and e-value cutoffs on the produced alignments.

The BLASTn search was repeated for the NCNS and the Ensembl datasets (**Table 3.1**) and the resulting ROC curves (**Figure 3.3 B,C**) confirmed the e-value and query alignment length as the best parameters to identify significantly conserved lncRNAs (AUC NCNS: e-value 0.76, alignment length 0.66; AUC Ensembl: e-value 0.82, alignment length 0.70). The identified cut-offs are as follows: NCNS) e-value 1e-04, alignment length 66; Ensembl) e-value 2e-04, alignment length 62. The results and the annotations of the homology searches for all 3 datasets can be found in the Additional file 2 of my publication (http://www.biomedcentral.com/content/supplementary/1471-2105-14-s7-s14-s2.xls).

Well characterised lncRNAs like the *HOTAIR* and the *Xist* are reported to contain short motifs which may form stem-loop structures to interact with protein

complexes, the retention of the structure deemed important for the lncRNA function (He et al., 2011b; Wutz, 2011). I wanted to test for the presence of structural conformations within the conserved sequences, which may lie undetected in a primary sequence alignment. Hence I compared the secondary structure of the aligned regions for the CNS and rCNS datasets, to test for RNA secondary structure constraint using the RNAz program (Washietl et al., 2005). The RNAz method detects conservation by comparing the sequence along with the predicted mRNA secondary structure of the aligned regions. The program frames two principal measures: 1) RNA secondary structure conservation and 2) thermodynamic stability. I have used three measures coming from the RNAz results to build the ROC curves: ratio of pairwise identity by sequence conservation index, Z score and P value (1/P value) (**Figure 3.4**). The sequence conservation index demonstrated a positive performance (AUC 0.74) in accordance with previous reports about structural conservation of conserved lncRNAs (McCutcheon and Eddy, 2003; Seemann et al., 2007). However, the performance of RNAz as a classifier is not as sensitive as BLAST e.value (AUC 0.74 vs 0.79) and RNAz alone cannot filter for low complexity regions in the alignments. Hence I decided to not consider the RNAz results further, in my analyses.

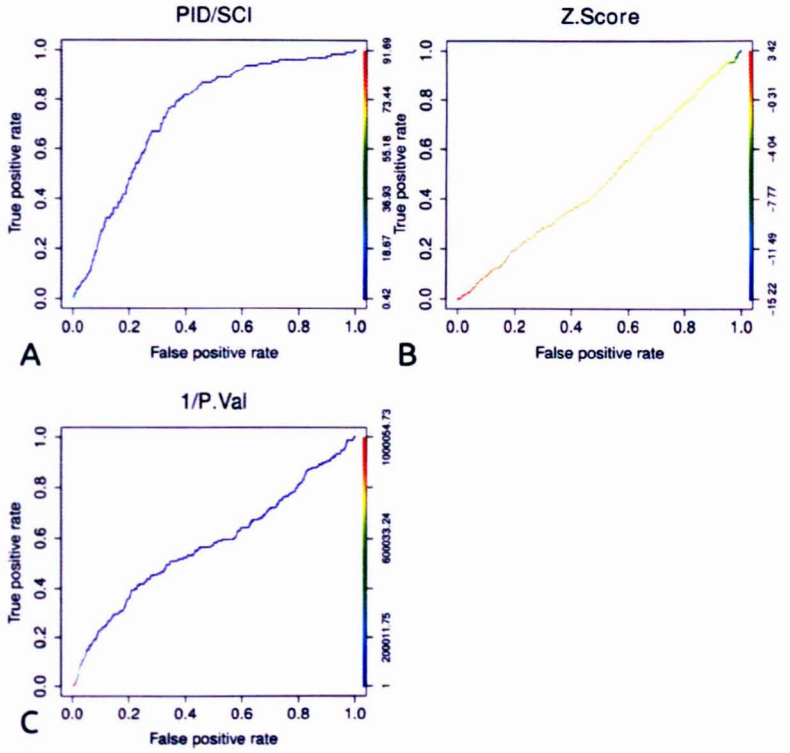


Figure 3.4 ROC curve for structural conservation of CNS lncRNAs dataset. **A)** Pairwise identity/Sequence conservation index (AUC 0.74), **B)** Z score (AUC 0.47) and **C)** inverse P-value (AUC 0.57) for the mouse CNS constrained lncRNAs against the zebrafish phastcons elements.

3.3.3 Comparison of the genomic contexts of mouse lncRNA and fish phastCons pairs predicted to be conserved

The position of the conserved regions with respect to other coding genes identifies those regions which do associate with an overlapping coding gene. Thus I mapped and compared each conserved element in the respective genic context of both analyzed organisms. The 11 putatively conserved lncRNAs in the CNS dataset showed homology to 10 phastCons elements. The NCNS dataset had 23 lncRNAs showing homology to 21 phastCons and the 250 conserved Ensembl lincRNAs showed homology to 209 fragments from 197 phastCons elements. I compared the

location of the conserved regions with annotated genes from the Ensembl database (**Table 3.2**). More than 30% of the conserved fragments in mouse and 60% in zebrafish, from the CNS dataset, overlap non-coding regions (intergenic, intronic or non-coding exon). The numbers increase for the NCNS dataset (mouse: 56%, zebrafish: 72%) but for the conserved Ensembl dataset only a minor fraction of elements overlap non-coding regions (mouse: 27%, zebrafish: 18%). The CNS and the NCNS lncRNAs are classified strictly based on their position (mainly intergenic) while in the Ensembl dataset, the candidate lncRNA fragments may overlap an external exon of a coding gene in the same chromosomal domain more frequently. However, these lncRNAs must still be considered non-coding because the orientation of the transcripts is in antisense to the protein coding genes they partially overlap. A well known example is the mouse *Xist* lncRNA (Ensembl gene ID: ENSMUSG00000086503) which overlaps a protein coding gene (exon to exon overlap). Antisense transcription (specially involving coding/non-coding pairs) is reported to occur genome-wide in unicellular organisms (Ni et al., 2010; Passalacqua et al., 2012), plants (Lu et al., 2012) and mammals (Conley and Jordan, 2012; Katayama et al., 2005). The antisense transcripts of mammalian genomes have been linked to the regulation of neighboring or overlapping protein-coding and small non-coding genes (Carrieri et al., 2012; Ebralidze et al., 2008; Hawkins and Morris, 2010).

Mouse					
Dataset	Total aligned regions	Coding exon overlap	Noncoding exon overlap	Intron overlap	Intergenic
CNS	11	7	0	3	1
NCNS	23	10	3	5	5
Ensembl	250	183	31	17	19
Zebrafish					
CNS	10	4	0	1	5
NCNS	21	4	2	3	12
Ensembl	209	171	6	9	23

Table 3.2 The genomic locations for the number of mouse lncRNA fragments and zebrafish phastCons regions found to be conserved. The location is deduced with respect to the coding region of the mouse and zebrafish genomes in the area of alignment.

A large proportion of antisense transcripts in humans belong to the class of long non-coding RNAs (Morris and Vogt, 2010) which can influence the expression of protein coding genes in *cis* as suggested in a previous report (Ponjavic et al., 2009). They are also reported to be associated with enhancers of neighboring coding genes in mouse neurons (Kim et al., 2010) and human (Ørom et al., 2010b). I chose to test if the function of flanking coding genes corroborates the functional conservation suggested for each mouse and zebrafish conserved non-coding pair. I identified the coding genes flanking and overlapping the conserved aligned regions in zebrafish and mouse, and evaluated their homology relationships. The search for orthologs was performed, scanning a window of 1 megabase flanking the conserved elements in either direction in the 2 genomes (see methods) (**Figure 3.5**). The **Figure 3.5** shows the percentages of conserved mouse lncRNAs sharing

orthologous coding gene in the corresponding zebrafish genomic context. All the lncRNA conserved fragments showed at least one ortholog pair from the CNS and the NCNS along with 80% of the Ensembl datasets supporting the hypothesis of syntenic conservation.

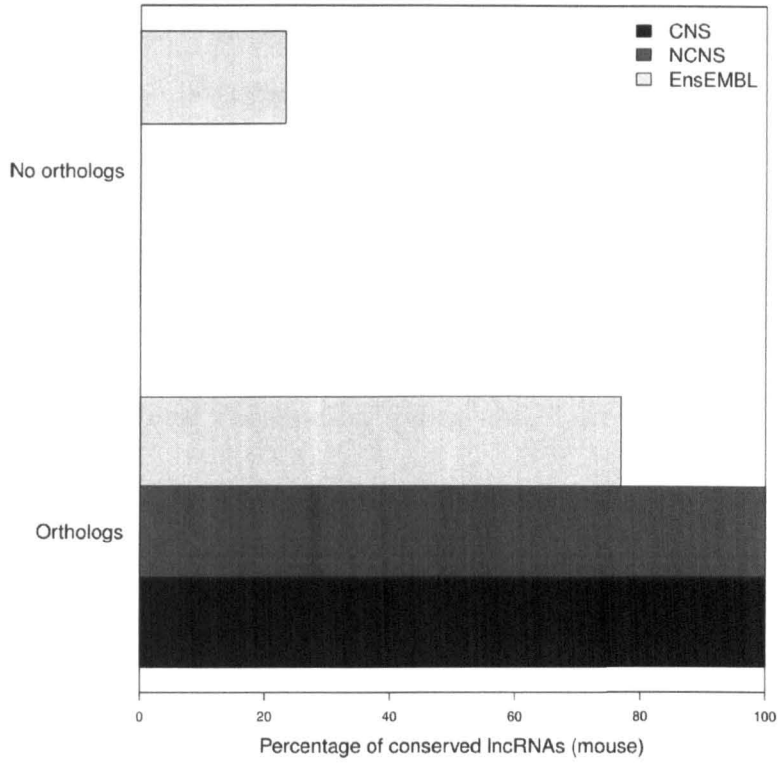


Figure 3.5 Orthologous protein coding genes flanking and/or overlapping conserved lncRNAs. The figure shows the percentage of mouse lncRNAs from CNS, NCNS and Ensembl datasets, conserved with a zebrafish phastCons element and sharing orthologous coding genes flanking or overlapping the region of conservation in zebrafish. The x-axis represents the percentage of conserved segments and the y-axis represents the group of conserved segments in different datasets, with or without synteny.

3.3.4 Functional enrichment analyses of the protein coding genes proximal to the conserved regions

Past reports have indicated a preference for lncRNAs to lie proximal to coding

genes with similar functionality, specifically genes involved in nervous system development and regulation of transcription (Guttman et al., 2009; Ponjavic et al., 2009). Further, based upon the similarity of the transcription specificity the lncRNAs are also suggested to be functionally related to the coding genes in their vicinity (Marques and Ponting, 2009; Ponjavic et al., 2009). Thus, in order to understand the potential biological role of the identified candidate lncRNA sequences I performed gene ontology and tissue specific expression enrichment analyses on the coding genes flanking the conserved fragments for the Ensembl dataset. The basic hypothesis behind the analysis was to check for a functional association between the long non-coding genes and their flanking coding genes. Significantly enriched GO biological process categories and tissue of expression for the coding genes flanking or overlapping conserved lncRNAs in zebrafish and mouse were considered. The analysis was performed using DAVID (Huang et al., 2009a, 2009b) at an EASE score cutoff of 0.05. The EASE score is a p-value adjustment method specifically designed for biological large-scale studies. It penalizes the significance of categories supported by few genes and favors more robust categories in respect to the Fisher exact probability. It is more conservative than the pure Fisher exact probability and less conservative than the Benjamini and Hochberg FDR (Hosack et al., 2003). The CNS and NCNS datasets were combined for the enrichment analysis to generate a dataset of reasonable dimensions to perform enrichments discovery (**Figure 3.6 A,B**) while the Ensembl dataset was analysed independently (**Figure 3.7 A,B**). The enriched GO terms for both the analyses included development, regulation of transcription and nucleic

acid metabolism as the major theme of functions in agreement with previous reports in mouse and zebrafish (Aprea et al., 2013; Guttman et al., 2010; Pauli et al., 2011a; Ulitsky et al., 2011). Tissue enrichment analyses were also performed to check if the selected genes showed an enrichment for being expressed in similar specific tissues. From this analysis neural and developmental related tissues resulted to be significantly enriched in both the species (**Figure 3.6 C,D; Figure 3.7 C,D**). These results are consistent with previous studies showing that lncRNAs play a fundamental role in regulation, neural development and plasticity (Mercer et al., 2008; Qureshi et al., 2010). The coding genes flanking conserved lncRNAs in mouse show a significant enrichment to be expressed in neural tissues yet other tissues like the lung and liver also feature prominently. This may indicate a possible sub-functionalisation of lncRNAs or a better representation of diverse set of tissue expression data in mouse. Taken together, these analyses highlight a conserved pattern of functions and expression domains of coding genes associated with conserved lncRNA fragments.

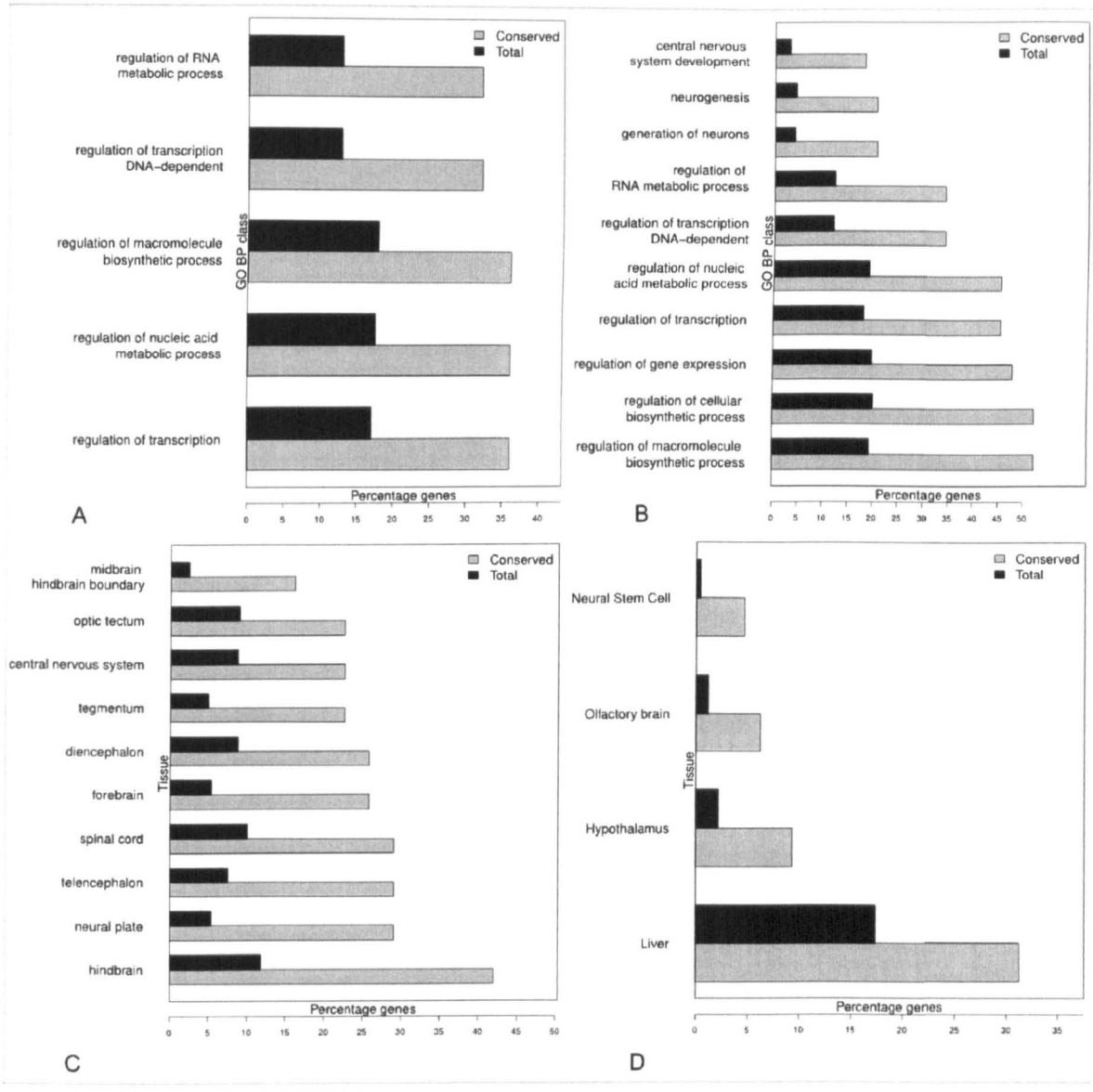


Figure 3.6 Function and expression of proteins flanking the conserved elements of the CNS and NCNS dataset. GO biological process term (level 5) enrichment of **A)** flanking proteins of conserved elements in zebrafish **B)** flanking proteins of conserved elements in mouse for the CNS and NCNS dataset. Tissue enrichment of **C)** flanking proteins of putative conserved elements in zebrafish **D)** flanking proteins of conserved elements in mouse for the CNS and NCNS dataset. A, B, C, D: GO terms and tissue of expression are listed only if they are significantly over-represented according to the EASE score. Grey bars indicate the percentages of genes associated to the respective functional classes from the group of genes flanking the identified conserved elements. Black bars indicate the percentages from the entire transcriptome of the given species

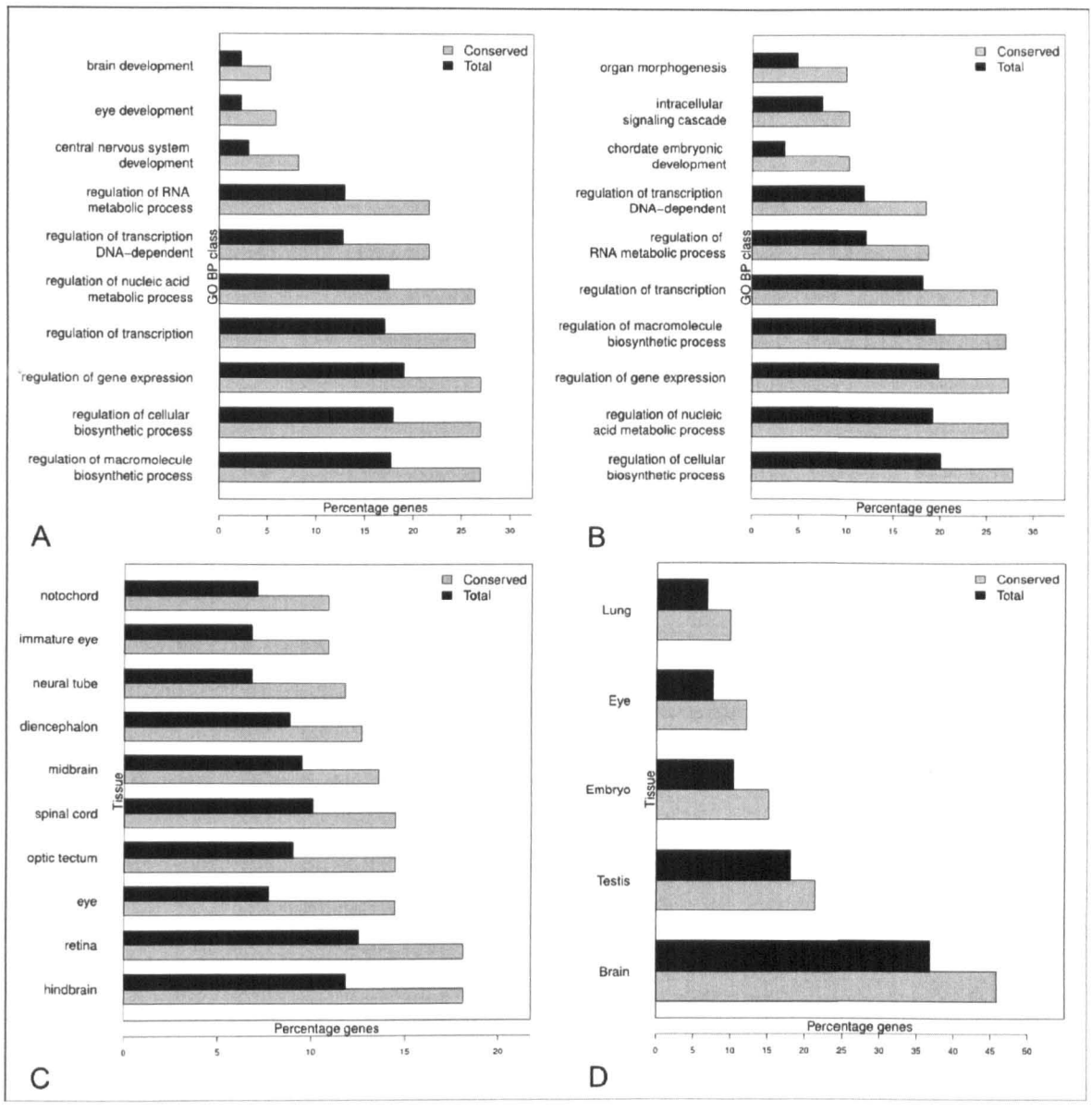


Figure 3.7 Function and expression of proteins flanking the conserved elements of the Ensembl dataset. GO biological process term (level 5) enrichment of **A)** flanking proteins of conserved elements in zebrafish **B)** flanking proteins of conserved elements in mouse for the Ensembl dataset. Tissue enrichment of **C)** flanking proteins of putative conserved elements in zebrafish **D)** flanking proteins of conserved elements in mouse for the Ensembl dataset. A, B, C, D: GO terms and tissue of expression are listed only if they are significantly over-represented according to the EASE score. The 10 top-scoring classes are present into the plots. Grey bars indicate the percentages of genes associated to the respective functional classes from the group of genes flanking the identified conserved elements. Black bars indicate the percentages from the entire transcriptome of the given species.

3.3.5 Overlap of conserved lncRNA segments with Ultra conserved elements

In the past, non-coding regions with high level of sequence conservation across vertebrates were reported in humans. These elements are known as Ultra Conserved Elements (UCEs) and show close to 100% sequence identity with mouse and many of them are conserved also in fishes. UCEs are greater than 200 nucleotides in length and observed to lie proximal to coding genes related to development, regulation of transcription (Bejerano et al., 2004) and cancer related loci (Calin et al., 2007). A small fraction of them overlap protein coding exon, however UCEs are mainly non-coding and intergenic in nature. Although a large fraction seems to be transcribed and/or to function as enhancer they do not overlap current collections of transcripts (Calin et al., 2007; Licastro et al., 2010; Pennacchio et al., 2006). In order to check if the identified sequences might belong to the ultraconserved family of elements I measured their overlap with UCEs reported in two previous studies (Bejerano et al., 2004; Sakuraba et al., 2008). In total four UCEs were found to overlap conserved regions from lncRNAs of the Ensembl dataset while a single lncRNA from the NCNS dataset showed overlap with a single UCE. Hence I concluded that the conserved regions identified in this study are not enriched for and do not correspond to UCEs elements.

3.3.6 Expression potential of conserved regions in zebrafish

The presence of expressed sequence tags (ESTs) overlapping the region of conservation argues for an active transcriptional output in the given region. It is important to note though, that experimental validation of selected conserved

regions in zebrafish is necessary to provide a conclusive evidence of lncRNA transcription. I checked for the overlap of zebrafish ESTs in the region of conservation. Around 60%, 45% and 70% of the predicted CNS, NCNS and Ensembl conserved regions overlap at least one EST in zebrafish. Interestingly, a random selection of ~1,200 non-repeated genomic regions, similar in size to the conserved regions, gave only 8% of overlap with ESTs (two sample proportion test: p-value: CNS 7.5e-08; NCNS 5.2e-09 and Ensembl 2.2e-50). The results suggest that the majority of the conserved regions fall under genomic regions being actively transcribed. Further, in order to support the transcriptional potential of the zebrafish conserved fragments I performed an overlap analysis with the recently published zebrafish candidate lncRNAs (1824 transcripts) resulting from RNAseq experiments (Pauli et al., 2011a; Ulitsky et al., 2011). The comparison of all the predicted conserved regions with the published lncRNAs resulted in 6% of the conserved regions showing overlap with at least one reported lncRNA. It is important to point out that no definitive estimation of the number of lncRNAs expressed in an organism is currently possible. Such uncertainty arises from the fact that non-coding RNAs are expressed at lower levels as compared to coding genes (Cabili et al., 2011; Guttman et al., 2010; Pauli et al., 2011a). Computational identification of lncRNA transcripts from next-generation sequencing data remains a *"work in progress"* in terms of mapping reads to the genome, assembly of new transcripts, definition of background noise and cut-off parameters. Hence, a lack of overlap does not signify an absence of transcribed elements in zebrafish, but may reflect undetected transcripts. In order to test this hypothesis I mapped the raw

reads from the study (Pauli et al., 2011a) (SRA accession: SRP009426) on the zebrafish genome and computed the overlap between the mapped reads and all the conserved fragments. Interestingly, more than 90% of the predicted conserved regions in the zebrafish genome show overlap with at least one mapped read while only 25% of a set of randomly chosen genomic regions overlap at least one read (two sample proportion test: p-value < 2.2e-50). Checking for regions with more than 1,000 reads overlap, we found that 20% of the conserved regions resulted positive while only 4% of random regions showed such an overlap (two sample proportion test: p-value = 1.2e-15; **Figure 3.8**).

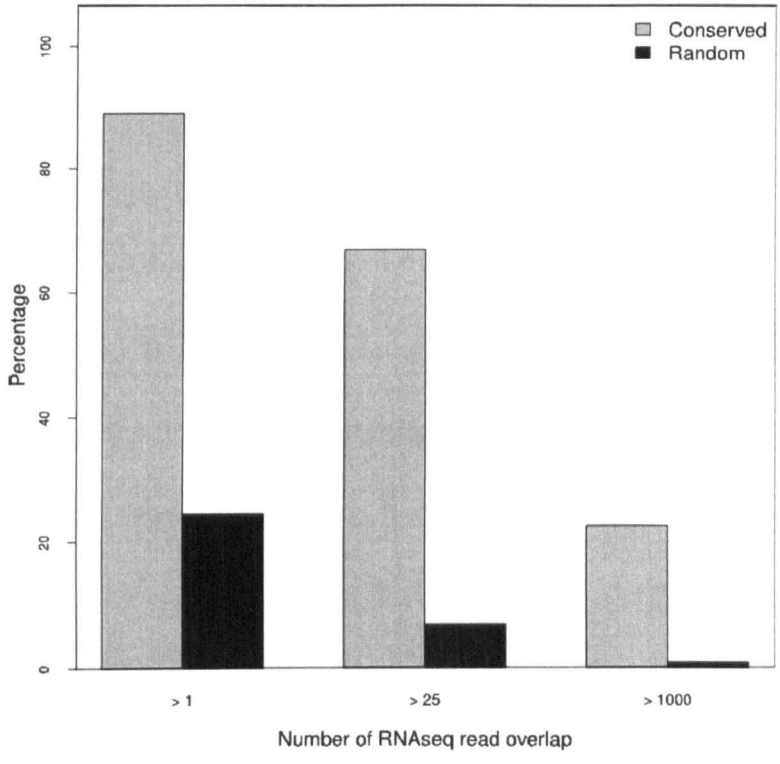


Figure 3.8 RNAseq data overlap on conserved zebrafish elements. The figure depicts the percentage of conserved elements in the zebrafish genome which show overlap with > 1, > 25 and > 1000 short reads (coming from RNAseq of zebrafish development stages) as compared against a set of random elements in the fish genome. The x-axis represents the number of overlapping sequencing reads and the y-axis represents the percentage of features with an overlapping read.

The highly significant differences between the conserved regions and the random sequences indicate that the RNAseq data supports transcriptional evidences in zebrafish for most of the regions predicted to be conserved lncRNAs. Finally I used tissue specific RNAseq data from another teleost fish (*Gasterosteus aculeatus*: stickleback) to extract information on the possible tissues where the conserved zebrafish regions might be expressed. I mapped the conserved zebrafish regions on the stickleback genome and then compared mapped regions with RNAseq reads from multiple tissues (heart, kidney, testis, liver, muscle, skin, gill, eye and brain). All conserved zebrafish regions mapped on the stickleback genome and ~85% of the regions had transcription support from the overlap of reads from at least one tissue (**Figure 3.9A**). The conserved regions in zebrafish corresponding to the CNS dataset show a higher expression level in the brain (**Figure 3.9B**), in concurrence with the expression pattern of the mouse CNS specific lncRNAs. In contrast the NCNS and Ensembl regions are expressed in different tissues at a more basal level (**Figure 3.9 C, D**). The results confirm the brain specific expression of the CNS conserved regions also in zebrafish while for the NCNS dataset no exact conclusion can be drawn due to lack of read coverage in the region of alignments. Yet the observation that the Ensembl and NCNS sequences show positive expression levels in several tissues partially accounts for the corresponding mouse transcript models beings assembled from multiple tissues (Flicek et al., 2011; Ponjavic et al., 2009). The above analysis showed the similarity of transcriptional domains between the conserved mouse lncRNAs and their corresponding zebrafish fragments in the CNS dataset. The availability of

sequencing datasets covering a diverse group of tissues at high resolution can further aid in understanding the similarity of the transcriptional outputs of the NCNS and Ensembl datasets.

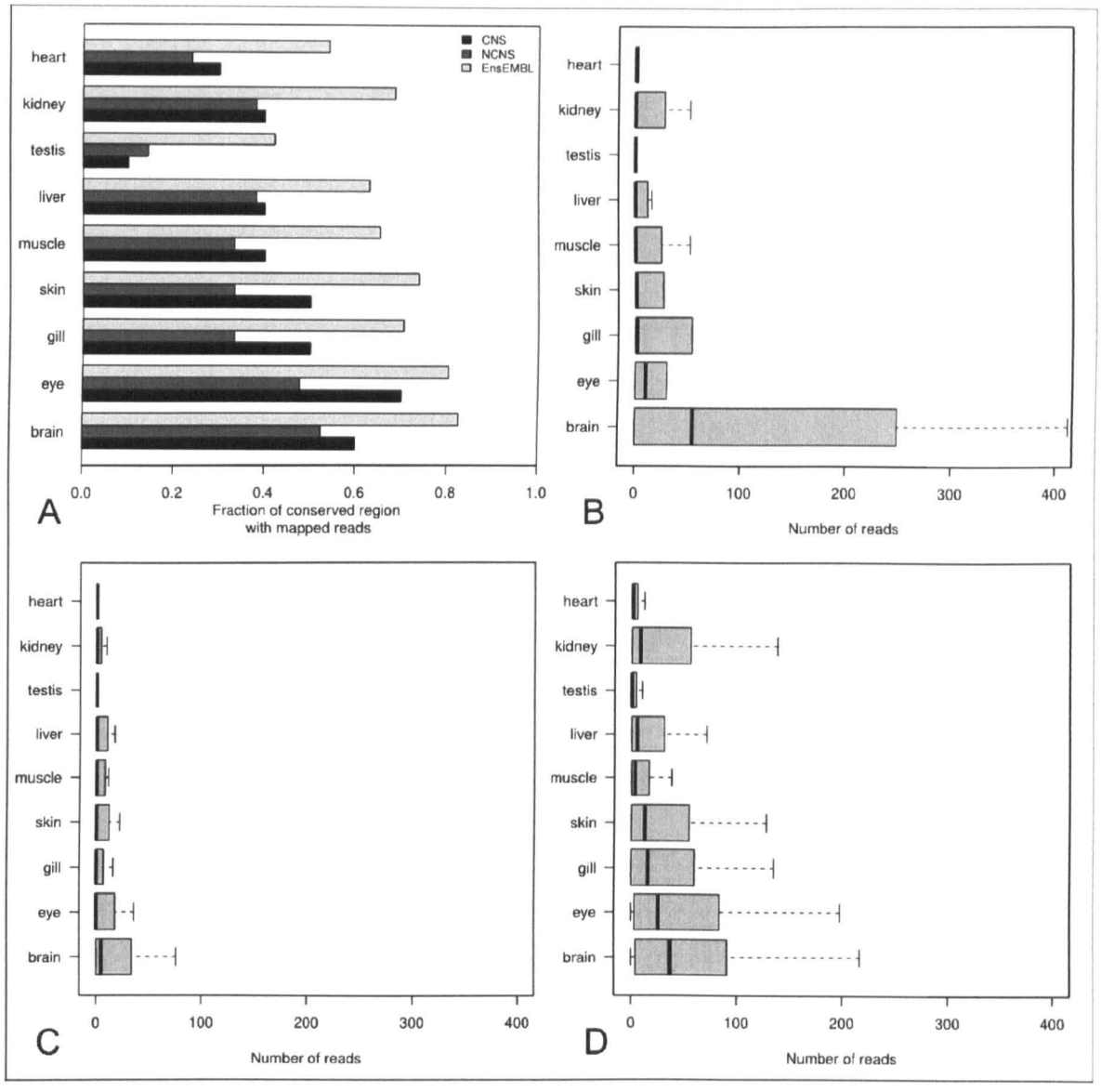


Figure 3.9 Tissue specific expression of conserved zebrafish regions mapped on the stickleback genome. **A)** Fraction of conserved regions in the CNS, NCNS, Ensembl datasets showing overlapping RNAseq reads from specific tissues of the stickleback **B)** Boxplot representing the number of reads from each tissue mapping on each conserved region coming from CNS dataset **C)** Boxplot representing the number of reads from each tissue mapping on each conserved region coming from the NCNS dataset **D)** Boxplot representing the number of reads from each tissue mapping on each conserved region coming from Ensembl dataset. Boxplots do not show outliers.

3.3.7 Examples of conserved lncRNAs

To better demonstrate the utility of my analysis I will discuss below an example each, of a conserved lncRNA region from the CNS, NCNS and Ensembl datsets. The first example is of a conserved element belonging to a cDNA sequence (*AK020962*) expressed in the mouse brain (**Figure 3.10 A**). The cDNA sequence comes from the CNS dataset and is classified as a lncRNA in a previous published study (Ponjavic et al., 2009). The cDNA sequence partially overlaps the UTR intron of the LIM Domain Only 3 (*LMO3*) coding gene. The corresponding conserved region in zebrafish is intergenic but flanked by the *LMO3* coding gene (**Figure 3.10 B**). The *LMO3* gene is known to be a transcriptional regulator (Hui et al., 2009) and is reported to be involved in cell proliferation and differentiation during embryonic development (Aoyama et al., 2005). It is also implicated in neuroblastoma through its interaction with the neuronal transcription factor *HEN2* (Aoyama et al., 2005). The zebrafish sequence shows a conservation of 96 base pairs with the murine lncRNA *AK020962* at an e-value of 4e-21 and 88% identity.

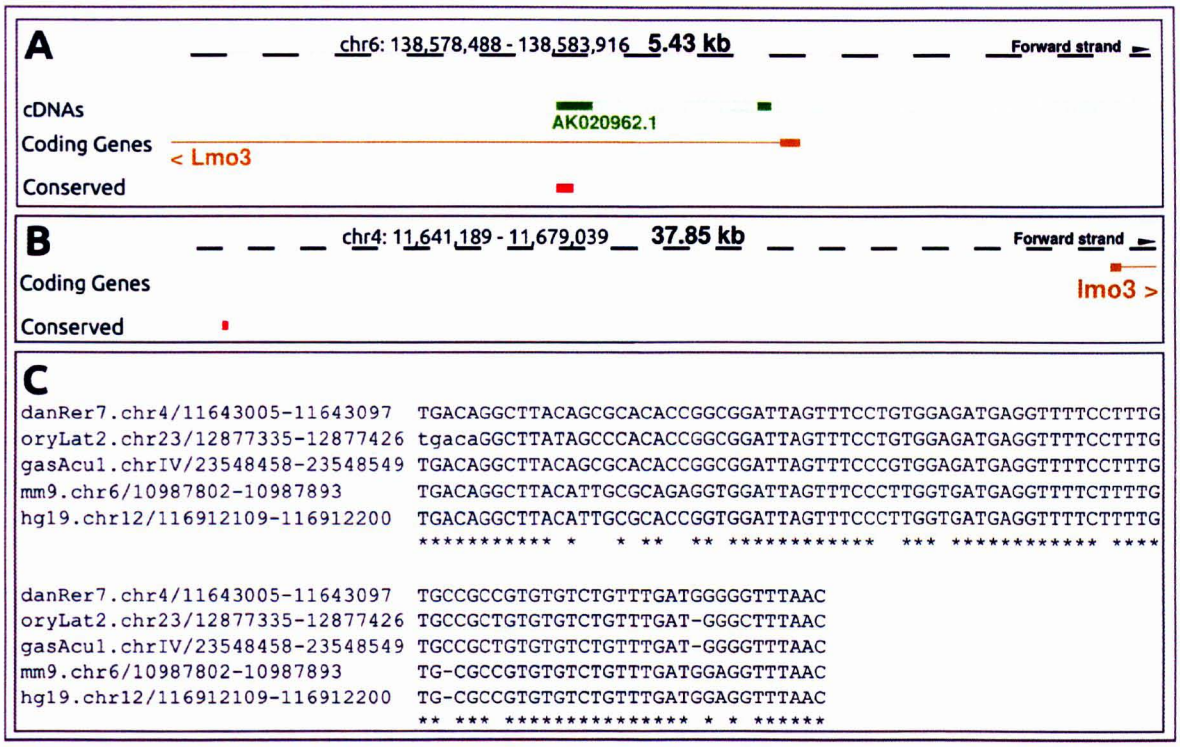


Figure 3.10 Genome browser screen-shots for a predicted conserved lncRNA **A)** The putative conserved lncRNA in the mouse genome. The red box indicates the conserved fragment. Green box above the conserved element mark the exons of the lncRNA *AK020962*. The conserved element lies inside an intron of the *LMO3* gene. **B)** Corresponding conserved region in the zebrafish genome indicated by the red box **C)** Section of multiz8way whole genome alignment (zebrafish as reference along with 7 vertebrates) overlapping the conserved region in zebrafish.

The second example comes from the NCNS dataset where the conserved region falls within a mouse cDNA sequence (*AK054275*) (**Figure 3.11 A**) classified as a lncRNA in a previous study (Ponjavic et al., 2009). The lncRNA falls in an intergenic region and belongs to the NCNS dataset. The lncRNA sequence lies close to the Ligand Dependent Nuclear Receptor Corepressor (*LCoR*) coding gene. The corresponding conserved region in zebrafish is also flanked by the *LCoR* gene and overlaps an intergenic EST sequence (CK360979) potentially non-coding in

nature (**Figure 3.11 B**).

Figure 3.11 Genome browser screen-shots for a predicted conserved lncRNA **A)** The putative conserved lncRNA in the mouse genome. The red box indicates the conserved fragment. Green box above the conserved element mark the exons of the lncRNA *AK054275*. The conserved element lies inside an intron of the *LCoR* gene. **B)** Corresponding conserved region in the zebrafish genome indicated by the red box. The purple box above the conserved region represents an overlapping EST sequence **C)** Section of multiz8way whole genome alignment (zebrafish as reference along with 7 vertebrates) overlapping the conserved region in zebrafish.

The zebrafish sequence shows a conservation of 75 base pairs with the mouse lncRNA *AK054275* at an e-value of 5e-09 and 84% identity. The *LCoR* gene is a transcriptional co-repressor and is expressed in diverse range of tissues (Fernandes et al., 2003). A recent report mentions a close coordination between the *LCoR*

protein with Kruppel-like factor 6 (*Klf6*) and Histone deacetylases (*Hdacs*) to regulate the expression of many target genes notably that of Cyclin dependent kinase inhibitor (*Cdkn1a*) (Calderon et al., 2012). The final example comes from the Ensembl dataset, a lncRNA predicted by the Ensembl lincRNA pipeline in the mouse genome (Flicek et al., 2011). The ID of this lncRNA has been updated in the current Ensembl version (Current ID: Gm26672; Past ID: Gm16882) (**Figure 3.12 A**). The region of conservation falls within the last exon of the lncRNA, the exon itself is completely imbricated in the intron of the Protocadherin Gamma Subfamily A, 9 (*Pcdha9*) gene. The corresponding conserved region in zebrafish lies in an intergenic region but proximal to the zebrafish *Pcdh2g16* gene (**Figure 3.12 B**). The zebrafish sequence shows a conservation of 77 base pairs with the mouse lncRNA *AK054275* at an e-value of 3e-09 and 83% identity. The protocadherins are a diverse group of cadherin protein expressed predominantly in the nervous system and implicated in cell recognition, cell signaling and development of neuronal circuits (Morishita and Yagi, 2007). Amongst the protocadherin family members, the gamma protocadherins (*Pcdhg*) are reported to be involved in synaptogenesis and apoptosis of interneurons in the developing spinal cord (Prasad and Weiner, 2011) as well as regulating the hypothalamic neuronal circuits to maintain energy homeostasis (Su et al., 2010) in mouse.

Figure 3.12 Genome browser screen-shots for a predicted conserved lncRNA **A)** The putative conserved lncRNA in the mouse genome. The red box indicates the conserved fragment. Grey box above the conserved element marks the exons of the lncRNA *Gm26672*. The conserved element lies inside an intron of the *Pcdhga9* gene **B)** Corresponding conserved region in the zebrafish genome indicated by the red box **C)** Section of multiz8way whole genome alignment (zebrafish as reference along with 7 vertebrates) overlapping the conserved region in zebrafish.

The examples discussed above aptly demonstrate the presence of sequence conservation in a select subset of mouse lncRNAs. The presence of the corresponding conserved regions in zebrafish near orthologous coding genes gives additional support to the predicted homology of the lncRNAs. Numerous prior reports have associated lncRNA expression and function with the development and differentiation of the nervous system. Indeed in examples discussed above I

find that the putative conserved lncRNAs in mouse and the corresponding conserved region in zebrafish lie near coding genes implicated in regulation of transcription and proliferation of neuronal circuitory. While an *in-depth* characterisation of the putative conserved candidates is required to establish the lncRNA function, the current results highlight the ability of my pipeline to predict candidate conserved lncRNAs which may play an important role in organism development and differentiation.

3.4 Conclusions

Unlike coding genes lncRNAs are devoid of a selective pressure to retain their nucleotidic sequence. Numerous past studies have emphasised on sequence homology being a poor metric to measure lncRNA conservation amongst species. Nonetheless I am able to demonstrate the presence of sequence conservation in a select set of mouse lncRNAs by comparing it with the conserved genomic regions of zebrafish. I demonstrate that between 4 and 11% of mouse lncRNAs (two constrained and a genome wide lncRNA dataset, 2,800 lncRNAs) are significantly conserved in zebrafish in agreement with the results by Ulitsky et al (Ulitsky et al., 2011) on a smaller dataset. Gene ontology analyses of protein-coding genes flanking the conserved elements, identified similar functional classes in both species to be significantly enriched, such as regulation of transcription and development. It is interesting to note that the coding genes, flanking the zebrafish homologs for mouse *CNS-specific* lncRNAs, were also enriched to be expressed in the brain. In order to detect sequence conservation, I have developed an analysis

pipeline which employs a sensitive procedure to systematically measure the homology of lncRNA sequences. The pipeline uses the widely accepted BLASTn program along with robust statistical analyses to define threshold values for identifying conservation. My analysis has predicted 4 to 11% of mouse lncRNAs to contain sequence blocks, conserved amongst vertebrates. It is important to note that the thresholds defined in my analysis result in a complete absence of false positives and majority of the predicted regions are not characterised in zebrafish as lncRNAs. This shows the ability of my pipeline to detect regions of conservation, which suggest the transcription of putative novel conserved lncRNAs. Further I found significant similarity between the expression domains and functional classes of the coding genes which beset the region of conservation in mouse and zebrafish. It is interesting to note that the subset of mouse lncRNAs expressed in the central nervous system have their corresponding zebrafish conserved regions and the flanking coding genes enriched to be expressed in neuronal tissues. Majority of predicted conserved regions in zebrafish are transcribed actively as evident from the overlap of ESTs and RNAseq reads. The putative conserved mouse lncRNAs provide a well annotated dataset to the community which is ideal to select interesting candidates for experimental validation. Finally I project this pipeline as an effective system to identify putative conserved lncRNAs in a diverse range of organisms. It is however important to point out that, coding genes, involved in functional mechanisms like organism development and regulation of transcription, are reported to co-localise with conserved non-coding sequences with a potential *cis*-regulatory function (Dermitzakis et al., 2002; Woolfe et al.,

2005). Thus, the discovered conserved lncRNAs might be enriched for such conserved *cis*-regulatory elements independently by the function of their transcript. To shed light on this issue, in the following chapter, I will specifically test the overlap of conserved non-coding elements with conserved regions in lncRNAs, estimating the enrichment of conserved gene-regulatory features to lie in the vicinity of positionally conserved long intergenic non-coding RNAs (lincRNAs).

Chapter 4

Conservation of microsynteny in vertebrate lincRNAs

4.1 Introduction

4.1.1 Retention of geneic order in coding and non-coding sequences

Long non-coding RNAs are predicted in diverse organisms but an effort to predict the putative functionally conserved candidates genome wide is still lacking. A primary reason is the paucity of primary sequence conservation in lncRNAs (Basu et al., 2013) and little knowledge about their secondary structure conservation (Novikova et al., 2012). An alternative strategy is to utilise the retention of orthologous flanking coding gene order as a possible mechanism to identify orthologous lncRNAs. In principle the approach looks simple, which is to identify lncRNAs in genomic loci with retained local order of coding genes. In practice the task is not trivial since whole genome analyses in metazoan species with low evolutionary turnover show a retention of chromosomal scale organization (macrosynteny) and a lack of conservation of local gene order (microsynteny) (Putnam et al., 2007, 2008; Srivastava et al., 2008, 2010). However, in particular cases evolution is known to favor microsynteny, exemplified by highly conserved non-coding elements (HCNEs) which regulate the expression of development and

differentiation related genes (Pennacchio et al., 2006; Shin et al., 2005; Woolfe et al., 2005). The HCNEs lie in clusters along with their target genes and other bystander genes maintaining a stretch of conserved gene order known as Genome Regulatory Blocks (GRBs) (Kikuta et al., 2007a), characterized by extensive microsynteny in metazoans (Engström et al., 2007; Kikuta et al., 2007a) as well as in plants (Baxter et al., 2012). The retention of microsynteny due to functional linkage between HCNEs and their target coding genes is more obvious in teleost fishes which have undergone whole genome duplication and rediploidisation to loose many bystander genes while keeping the HCNE-target gene association intact (Becker and Lenhard, 2007; Kikuta et al., 2007b).

4.1.2 Conservation of microsynteny in long non-coding RNAs

Although conserved non coding elements are associated with retention of local gene order due to long/short range cis-regulatory constraints there is little evidence to justify a similar claim for lncRNAs. A recent report detected 196 unique orthologous pairs (4% of human lincRNAs) of conserved long intergenic non-coding RNAs (lincRNAs) in human and mouse based upon genomic alignments of lincRNA loci (Managadze et al., 2013). I have shown previously that based upon conservation of sequence a similar percentage of mouse lncRNAs (4-11%) have putative orthologs in zebrafish (Basu et al., 2013). The results of my analysis were supported by another study which reported around 4% of zebrafish lncRNAs to show sequence conservation with their mammalian counterparts (Ulitsky et al., 2011). In fact the same study predicted a higher percentage of

zebrafish lncRNAs (~ 20%) enriched to lie near coding genes whose orthologs in human and mouse had an adjacent lincRNA (Ulitsky et al., 2011). This observation provided an impetus to probe deeper into the positional conservation of lincRNAs specially within vertebrate genomes where the local gene order is better conserved and there exist published lncRNA datasets (Derrien et al., 2012; Flicek et al., 2012b; Pauli et al., 2011a; Ulitsky et al., 2011). A question of primary importance is whether lncRNAs may associate with their genomic neighborhood over long evolutionary distances, due to an existing regulatory constraint. Hence I decided to develop a computational pipeline which can predict candidate long intergenic non-coding RNAs (lincRNAs) which retain their position between a chosen pair of species. I wanted to use the pipeline to identify a candidate set of lincRNAs which are predicted to be microsyntenic in human, mouse and zebrafish. I wanted to test whether the predicted microsyntenic lincRNAs lie enriched in their position in comparison to random locations in the genome. Further I wanted to check whether the predicted microsyntenic lincRNAs are under the influence of known non-coding regulatory features. Finally I intended to use gene expression, sequence conservation, abundance of regulatory features and chromatin interaction as measures to indicate a cis-regulatory constraint within the microsyntenic lincRNAs with respect to the associated coding genes.

4.2 Materials and Methods

4.2.1 Data sources

The long non-coding RNA dataset for human was downloaded from Gencode v17 (Harrow et al., 2012). All long non-coding RNAs predicted by the Ensembl pipeline (database version 72) were considered for mouse (Flicek et al., 2012b) and lncRNAs predicted by two prior published studies (Pauli et al., 2011a; Ulitsky et al., 2011) along with those classified by Ensembl (database version 72) (Flicek et al., 2012b) were pooled together for zebrafish. The genomic coordinates of coding genes and their homology relationships for each organism were downloaded from the Ensembl Compara database (Vilella et al., 2009) (version 72). The data retrieval from the Ensembl database was carried out using the Bioconductor (Gentleman et al., 2004) package biomaRt (Durinck et al., 2005). The chromosomal location of GRBs were obtained from the UCNEbase (Dimitrieva and Bucher, 2012). Histone monomethylation (H3K4Me1) and acetylation (H3K4Me3) marks for human and mouse embryonic stem cells were downloaded from the UCSC genome database and from a prior published study in zebrafish (Bogdanovic et al., 2012). The data sources for human and mouse are indicated below

- Human (Ram et al., 2011)
  - http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescH3k4me1StdPk.broadPeak.gz
  - http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescH3k27acStdPk.broadPeak.gz
  -
- Mouse (Ram et al., 2011)

- http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/wgEncodeLicrHistoneEse14H3k04me1ME0129olaStdPk.broadPeak.gz
- http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/wgEncodeLicrHistoneEse14H3k27acME0129olaStdPk.broadPeak.gz
- http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/wgEncodeLicrHistoneEsb4H3k4me1ME0C57bl6StdPk.broadPeak.gz
- http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/wgEncodeLicrHistoneEsb4H3k27acME0C57bl6StdPk.broadPeak.gz

The H3K4me1 peaks which overlapped an H3K27ac peak were considered active enhancers. The mean phastCons conservation score for all active enhancers was calculated and those with the top 25% scores were considered as conserved active enhancers. Insulator marks or *CTCF* binding sites for human and mouse were obtained from published studies (Bao et al., 2008; Nitzsche et al., 2011). Genome wide phastCons sequence conservation score for all the species were downloaded in WIG and BigWig format from the UCSC database. The files contain the genome wide conservation score of each base pair of the human genome. The conservation score is generated by the phastCons program (Pollard et al., 2010) using a hidden markov model method on whole genome alignments of mutliple species. The files downloaded are as follows.

- Human: Contains phastCons conservation score of each base in the human genome aligned to 45 other vertebrate species which include other mammals, birds, marsupials, reptiles and amphibians (http://hgdownload.soe.ucsc.edu/ goldenPath/hg19/phastCons46way/vertebrate/).
- Mouse: Contains phastCons conservation score of each base in the mouse genome aligned to 59 other vertebrate species which include

other mammals, birds, marsupials, reptiles and amphibians (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/phastCons60way/mm10.60way.phastCons60wayPlacental.bw).

- Zebrafish: Contains phastCons conservation score of each base in the zebrafish genome aligned to 7 other vertebrate species which include human, mouse, *Xenopus tropicalis*, fugu, medaka, stickleback, *Tetraodon* (http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/phastCons8way/vertebrate.phastCons8way.bw).

The conservation scores for each human chromosome in WIG format were converted to BigWig and merged using the wigToBigWig and bigWigMerge binaries (http://hgdownload.cse.ucsc.edu/admin/exe/). Transcription start sites across twelve developmental stages in zebrafish was obtained from a previous study (Nepal et al., 2013). The tissue expression data for 12 human tissues from Illumina Body Map was downloaded from the Ensembl database in BAM format (ftp://ftp.Ensembl.org/pub/release-73/bam/homo_sapiens/genebuild/) and for 25 tissues in mouse from the UCSC genome database also in BAM format (http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCshlLongRnaSeq/). Hi-C and ChIA-PET chromatin interaction data were downloaded from previous published studies (DeMare et al., 2013; Li et al., 2012a). Custom Perl scripts were used to map the chromatin interactions between different genomic loci.

4.2.2 SynLinc pipeline

The SynLinc pipeline was developed in the Perl programming environment (v5.12) and connects to a MySQL database (> 5.1.6) to store and retrieve information. The pipeline is comprised of three principal scripts (**Figure 4.1**) i) Build DB: This script

takes as input coordinates of coding and long non-coding genes/transcripts in BED6 format and uploads them into the database. ii) Build Homology: This script needs a file of pairwise orthologous gene IDs between two species to upload in the database and build the homology maps. iii) Build Synteny: This script analyses the data uploaded by the previous scripts to identify putative microsyntenic lincRNAs and gives a tabular output. The three scripts are discussed in detail below.

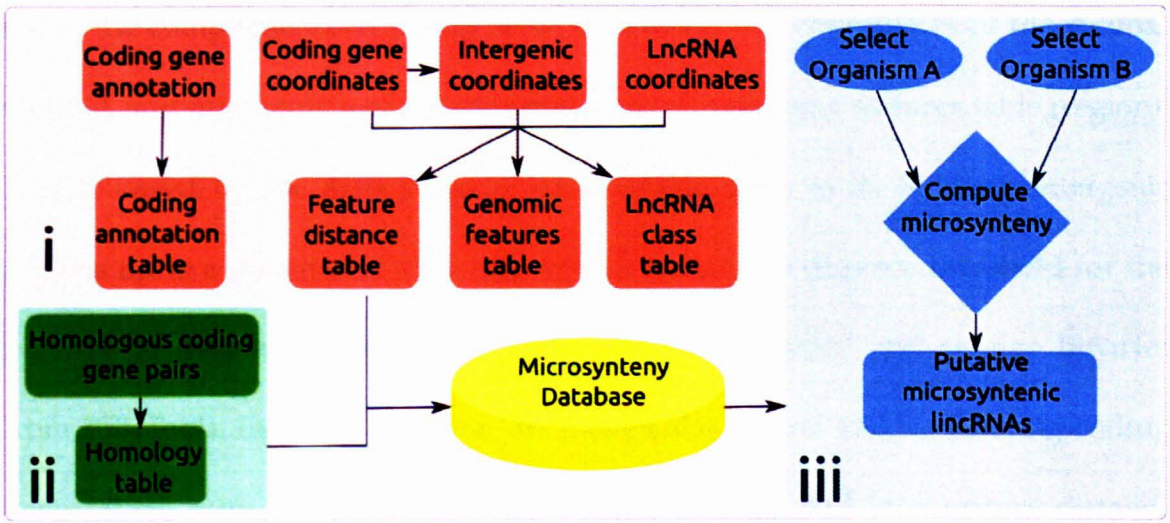


Figure 4.1 Schematic overview of the SynLinc pipeline. **i)** The Build DB script accepts coding and lncRNA gene coordinates along with gene annotation information of coding genes to populate various tables in the database. **ii)** The Build Homology script accepts pairs of orthologous/paralogous genes to populate a homology table. **iii)** The Build Synteny script takes as input names of two organisms whose data is already formatted and uploaded in the database and predicts putative microsyntenic lincRNAs between the two species.

4.2.2.1 Build database: Upload coordinates of coding and long non-coding RNAs into the microsynteny database

This script requires as input the genomic locations of all coding genes and long non-coding RNAs in BED6 format, a file containing coding gene identifier, symbol

and description and a file with chromosome names and sizes for a given organism. For each dataset the script needs an organism and data source name to be provided by the user (for example human and Ensembl72). The script requires prior installation of the BEDTools suite of programs (> v2.17) (Quinlan and Hall, 2010). It connects to a MySQL database to store the parsed data into respective tables. The coordinates of intergenic regions are calculated for a given dataset using the complementBed binary from BEDTools. The coordinates of the coding, lncRNA and intergenic regions are uploaed in the genomic features table (region). It is followed by the association of each coding gene to its flanking intergenic regions up to a distance of 1MB (the default maximum distance threshold for the pipeline to measure microsynteny) using the windowBed and overlap binaries from BEDTools. Information about each intergenic region and its flanking coding genes along with the distance of separation are uploaded in a feature distance table (fdist). Finally each lncRNA of a given organism is classified according to its position i) Intergenic: no overlap with a coding gene ii) Overlap: partial overlap with a coding gene iii) Containing: completely encompassing a coding gene iv) Contained: completely encompassed by a coding gene. The information about lncRNA classes is uploaded into the lncRNA class table (lnctype).

4.2.2.2 Build Homology: Upload pre-mapped gene identifiers predicted to be homologous between two species into the microsynteny database

The build homology script takes as input a tab delimited file of orthologous coding gene pairs between two organisms. The script searches the genomic feature

table for gene identifiers of each orthologous pair. If both genes are present in the genomic feature table the orthology information is uploaded in the homology table (protmap).

4.2.2.3 Build Synteny: Predict putative microsyntenic lincRNAs between two species in tabulated format

This script is the principal script of the pipeline and predicts putative microsyntenic lncRNAs between two organisms. It relies on the datasets uploaded using the Build DB and Build Homology scripts. The script only looks for "intergenic" class of lncRNAs. The basic working unit of the script are intergenic regions which contain a lincRNA (lincIGs). When two species are chosen to be analyzed the pipeline identifies all the orthologs of the coding genes lying near lincIGs, up to a given user specified distance, for species A in species B. If one of the orthologs lies near a lincIG in species B too, at the same specified distance threshold, the lincIG is deemed to be microsyntenic. Further, information on all possible combinations of lincRNAs falling in the predicted microsyntenic lincIGs are stored in a Perl hash object. This object is designed to be optimal for storing microsynteny information. It provides the ability to compute complex positional information quickly without redundancy to generate a tabular output. For example lincRNA_A is upstream to proximal coding genes P1, P2, P3 and downstream to P4, P5. The corresponding homologs of these proteins are P1`, P2`, P3`, P4`, P5`. LincRNA_A` is upstream to proximal coding genes P1`, P5`, P2` and downstream of P3`, P4`. The script builds a data structure for lncRNA_A to map all its predicted

homologs (**Figure 4.2**). The possible associations generated in the example are:

- UU:DD (P1:P1'-P4:P4') meaning that LincRNA_A and its putative ortholog lincRNA_A' lie upstream to P1 and P1' and downstream to P4 and P4'.
- DU:UD (P5:P5'-P3:P3')
- UU:UD (P1:P1'-P3:P3')
- DU:DD (P5:P5'-P4:P4')

Upstream and downstream (U and D) indicators permit to store the relative arrangement of the lincRNA with respect to each flanking protein coding gene taking into account strand information for both the elements of a coding/lincRNA pair. This strategy allows the script to classify the relative orientation that a lincRNA shares with its proximal coding genes according to the following orientation classes: convergent (tail to tail), divergent (head to head) or co-linear (same strand).

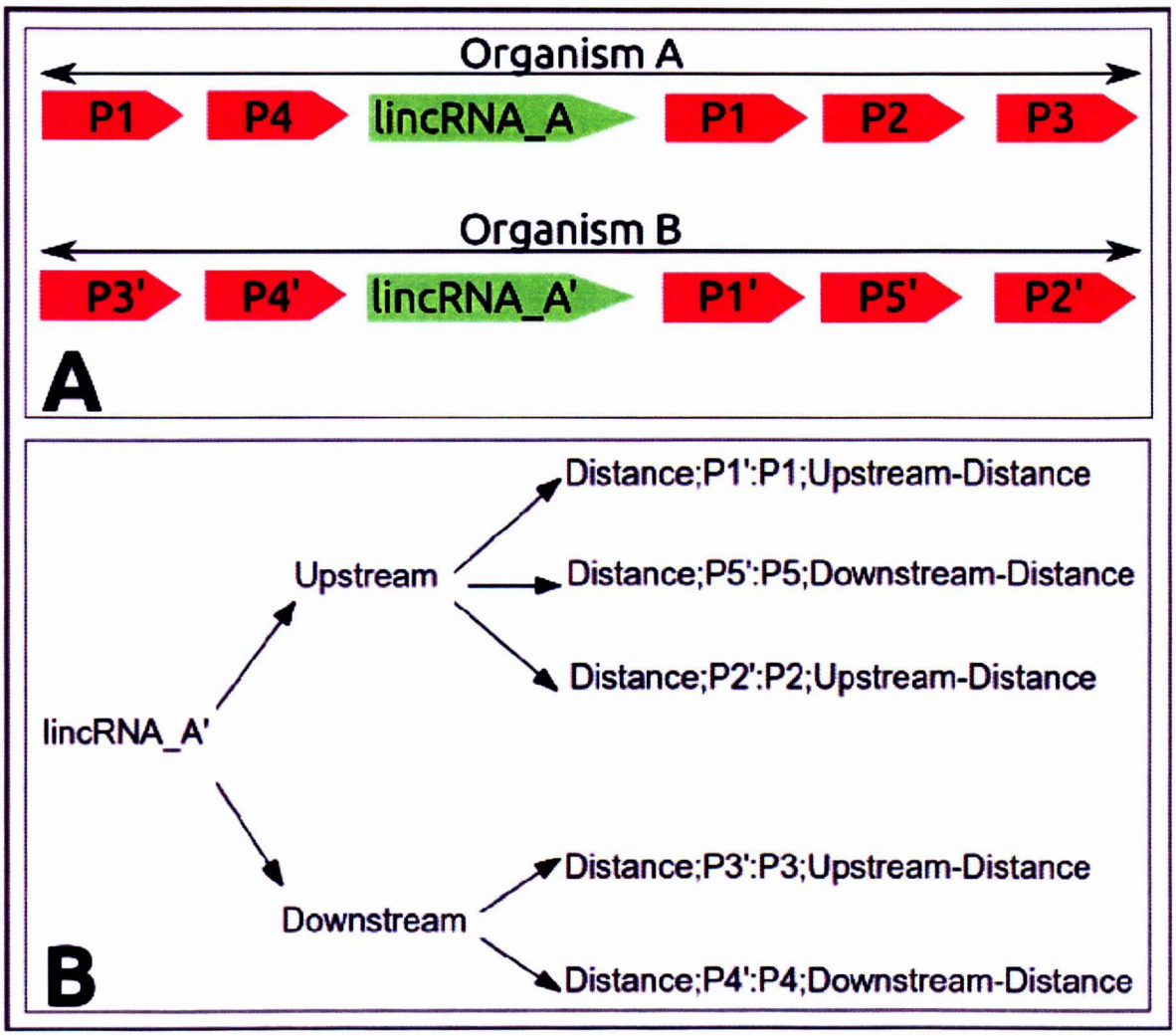


Figure 4.2 Structure of a Perl hash object used to store microsynteny information **A)** Organisation of orthologous coding genes flanking lincRNAs in two species. **B)** Representation of the organisation in terms of a perl hash object used to define microsyntenic associations in the SynLinc pipeline.

The script outputs a tab delimited text file which can be filtered by different parameters relating a coding gene with a lincRNA, like orientation, distance and symbol of the coding gene. The pipeline was run with a distance threshold of 1 base to identify putative microsyntenic lincRNAs between human-mouse, mouse-zebrafish and human-zebrafish. The lincRNAs which share at least one orthologus proximal coding gene in all the three species were categorized as vertebrate

microsyntenic lincRNAs (VMLs).

4.2.3 Computational characterisation of vertebrate microsyntenic lincRNAs

To check the significance of microsyntenic associations the pipeline was simulated 1000 times on each pair of organisms after random shuffling of transcript coordinates. During each shuffle the location of the coding and intergenic regions were kept constant while the coordinates of the lncRNAs were shuffled on the genome without any preference to overlap a coding gene or intergenic region. Shuffling of coordinates was done by using the shuffleBED binary from BEDTools. PhastCons conservation scores in biwig format were compared against genomic intervals between lincRNAs and their proximal coding genes using bigWigSummary utility from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/). Overlap and proximity of various genomic features with lincIGs and lincRNAs were calculated using the intersectBED and closestBED binaries from BEDTools. The multiBamCov script from BEDTools was used to obtain the count of reads for lincRNAs and coding genes across various tissues in human and mouse. An in-house R script was used to obtain overlap of CAGE peaks in the promoter regions of zebrafish coding genes and lincRNAs.

4.3 Results and Discussion

4.3.1 Association of lincRNAs with Genome Regulatory Blocks (GRBs)

The principal aim towards identification of microsyntenic lincRNAs is to associate them with a potential independent constraint which allows the retention of local

gene order. However, maintenance of local gene order is one of the core tenets of GRBs. The presence of microsyntenic lincRNA within a GRB may signify the influence of the GRB in establishing positional conservation rather than the lincRNA itself. Hence prior to running the SynLinc pipeline I checked whether lincRNAs extensively share chromosomal domains with GRBs by looking for the overlap of intergenic regions containing lincRNAs (lincIGs) with GRBs (**Figure 4.3**). I found that more than 90% of vertebrate lincIGs do not overlap GRBs, but a random subset of lincIGs have a higher propensity to overlap a GRB when compared to a random set of intergenic regions without lincRNAs (wlincIGs) in all the candidate species (5-10% for random lincIGs vs 2-4% for random wlincIGs; two sample proportion test, p-value < 2e-16).

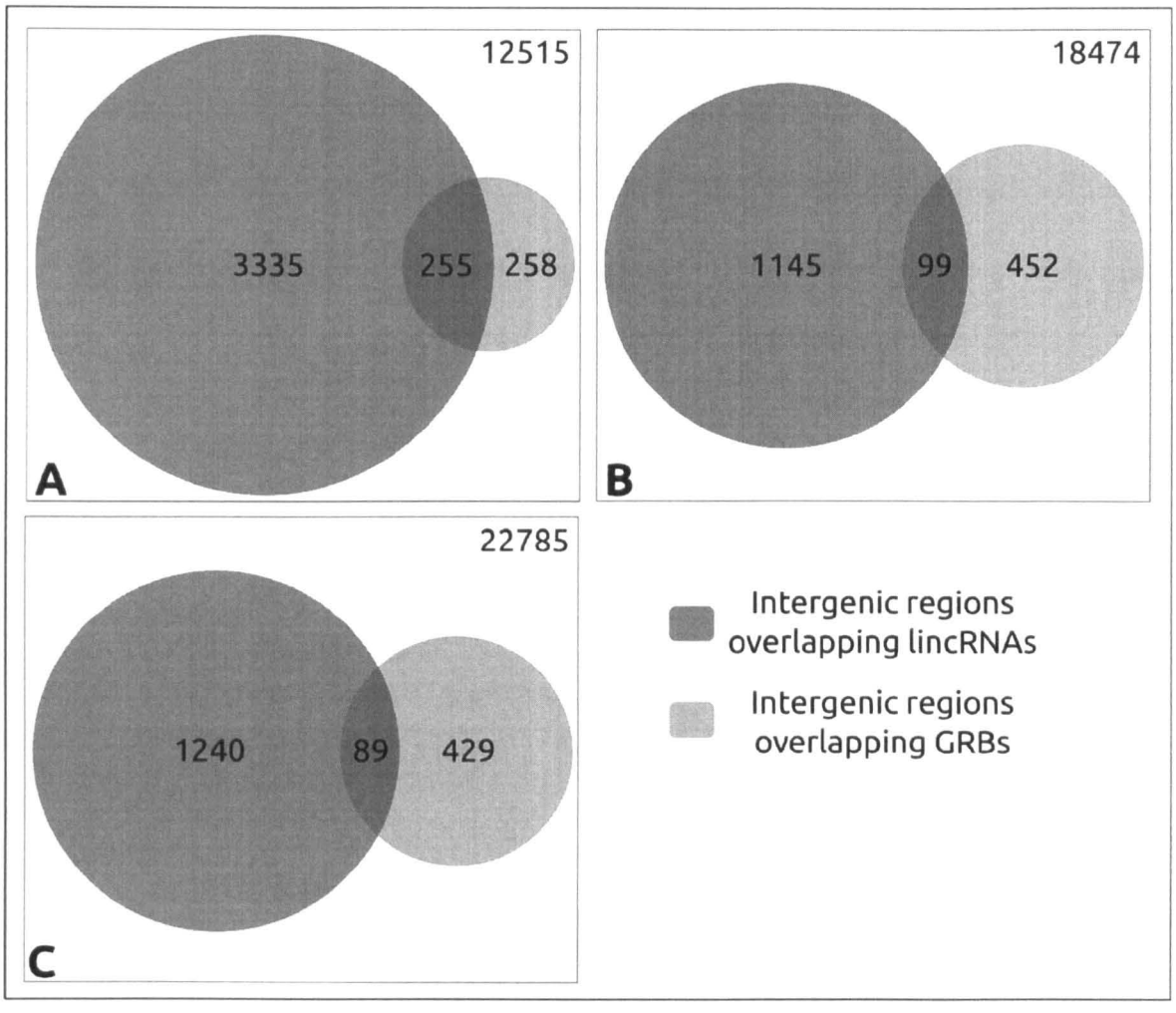


Figure 4.3 Overlap of intergenic regions containing lincRNAs with those overlapping Genome Regulatory Blocks in **A)** Human **B)** Mouse **C)** Zebrafish.

4.3.2 Prediction of Vertebrate Microsyntenic LincRNAs (VMLs)

The results suggests that a significant number of lincRNAs overlap GRBs than what can be explained by random chance and the GRB elements may influence the localisation of a lincRNA subset. Yet this number is not very large (5-10%) hence many lincRNAs may function beyond the influence of a GRB and large syntenic blocks, this is the reason why I developed the SynLinc pipeline that mainly checks for short microsyntenic blocks containing lincRNAs. The pipeline identifies

intergenic regions with lincRNAs in a given pair of organisms. A lincRNA is predicted to be microsyntenic if orthologous coding genes in both the organisms of choice contain a flanking lincRNA (**Figure 4.4**). The SynLinc pipeline was run on the human, mouse and zebrafish genomes to predict putative vertebrate microsyntenic lincRNAs (VMLs).

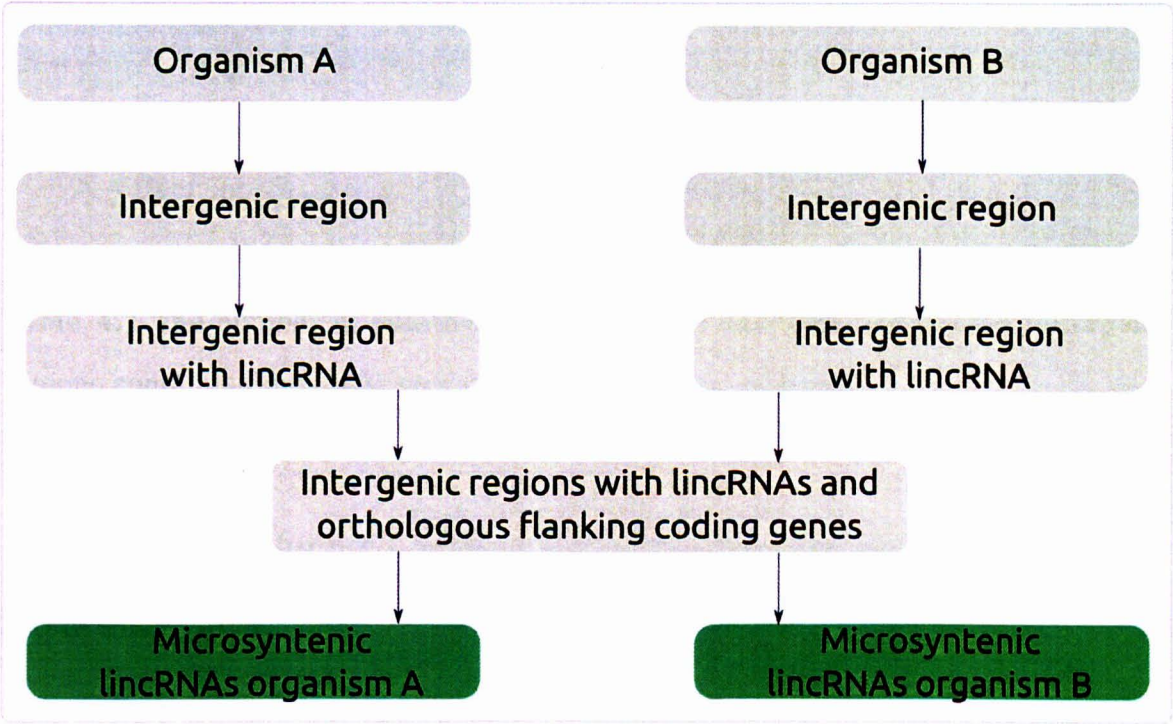


Figure 4.4 Workflow of the SynLinc pipeline for identification of microsyntenic lincRNAs between two organisms. The pipeline extracts intergenic regions in a given pair of organisms which contain lincRNAs (lincIGs). The lincIGs flanked by coding genes orthologous in both organisms are further selected as microsyntenic lincIGs. The selection of flanking coding genes depends upon a user specified distance threshold, of the coding gene from the closest end of a lincIG. The lincRNAs present inside microsyntenic lincIGs form the initial putative microsyntenic lincRNA dataset.

The pipeline predicted more than 200 VMLs in human, mouse and zebrafish (**Table 4.1**) based on the homology of the closest flanking coding genes for each

lincRNA. To consider association to only the closest flanking coding genes is a conservative approach which arises from the aim to reduce the number of predicted false negatives by solely considering coding/non-coding pairs which remain linked closely across evolution.

	Human	Mouse	Zebrafish
Total lncRNAs	13333	3982	3478
Total lincRNAs	7200	1721	2030
Vertebrate Microsyntenic lincRNAs	763	353	336
Total lincIGs	3590	1244	1329
Vertebrate microsyntenic lincIGs	261	206	209
LncRNA data source	Gencode17	Ensembl72	Ensembl72; Pauli *et al*; Ulitsky *et al*

Table 4.1 The number of putatively microsyntenic lincRNAs and lincIGs (intergenic regions containing lincRNA) predicted by the SynLinc pipeline.

Before proceeding further I define a few terms which are used frequently in the text of this chapter.

- VMLs: Vertebrate microsyntenic lincRNAs.
- VMLRs: A subset of VMLs which along with the position also retain their orientation with respect to a flanking coding gene.
- CCG: The closest coding gene for a lincRNA. In case of VMLs and VMLRs it is the closest coding gene which is used to define the microsynteny.
- VMLIGs: An intergenic regions which contains a VML.

To test the significance of the results random genomic segments were selected in the human, mouse and zebrafish genomes (size-matched to lncRNAs in each species). The SynLinc pipeline was run on the random genomic segments to

calculate the percentage of segments which are intergenic and microsyntenic across the three species after each randomization (**Figure 4.5**). This process was repeated one thousand times, each repetition considering the same set of organisms (human, mouse, zebrafish) with new random genomic segment datasets. The aim of the randomization is to demonstrate that the presence of a lincRNA near a coding gene, orthologous in human, mouse and zebrafish does not occur by random chance. The mean percentage of microsyntenic intergenic regions across randomized replicates was found to be significantly lower than those of lncRNAs (two sample proportion test, p-value Human: 10.08e-48, Mouse: 2.39e-94, Zebrafish: 1.72e-87).

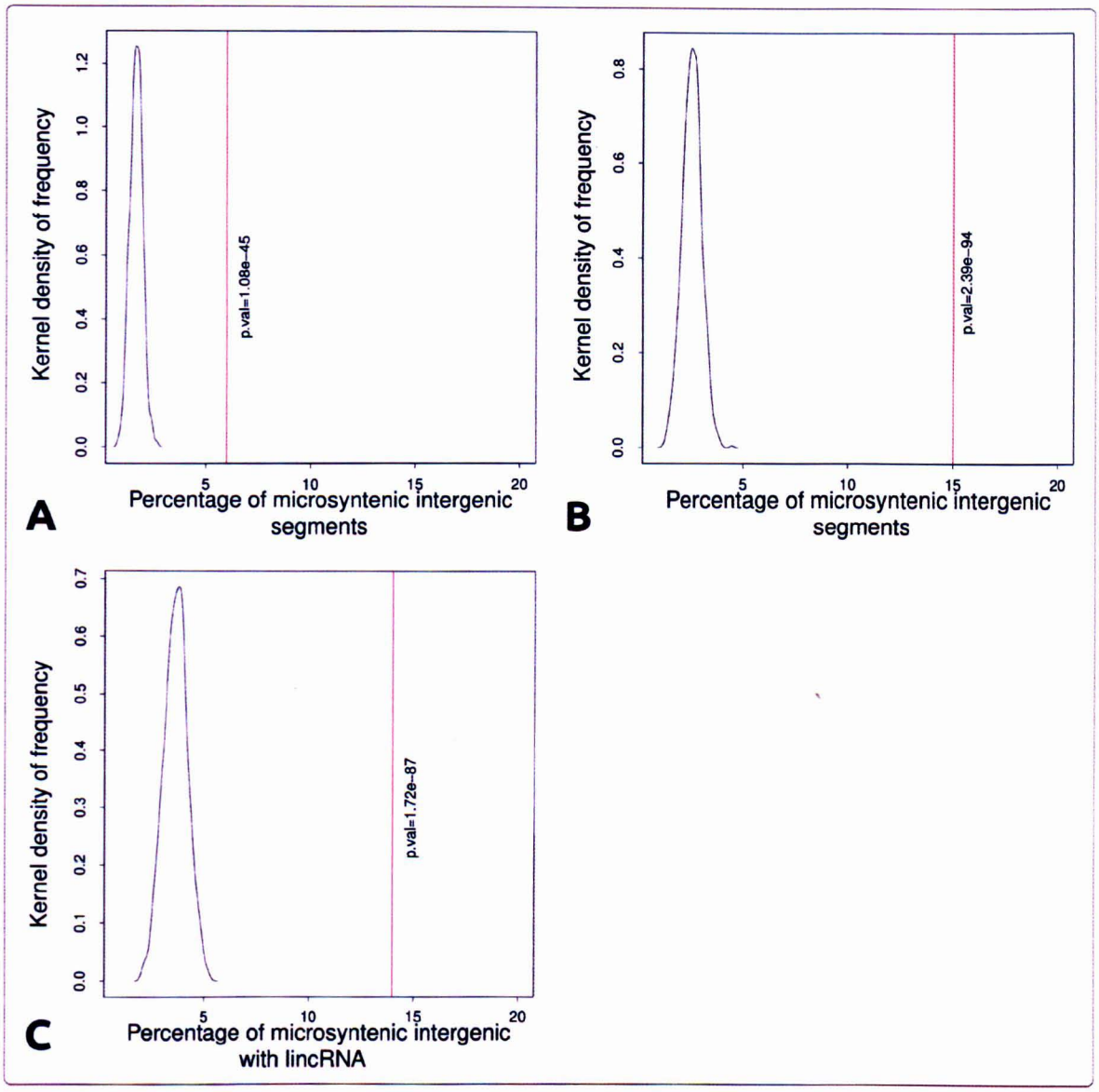


Figure 4.5 Frequency distribution of microsyntenic percentage in 1000 datasets of random genomic segments (size-matched to lncRNAs). Microsyntenic percentage signifies the percentage of random genomic regions for each dataset predicted to be intergenic and microsyntenic in human, mouse and zebrafish. The percentage distribution of plotted separately for **A)** human **B)** mouse **C)** zebrafish. The red line indicates the percentage of vertebrate microsyntenic lincRNAs in each species. The x-axis represents the percentage and the y-axis represent the frequency distribution.

To check for the proximity of VMLs to GRBs, I again compared the overlap between intergenic regions containing VMLs (VMLIGs) and GRBs. The aim was to

understand if the presence of GRBs has influenced the retention of the VMLs across evolution. Majority of VMLIGs lie outside GRBs (> 80%) in all the three species but the average distance of VMLIGs from a GRB is smaller than that of all lincIGs (two sample t test, p-values: Human 0.03; Mouse 0.04; Zebrafish 0.01) (**Figure 4.6**). However, in contrast to human and mouse, VMLIGs are not enriched to lie closer to GRBs in comparison to a random integenic region in zebrafish. Thus the results are not able to support the hypothesis of VMLs being enriched to lie closer to GRBs in zebrafish. This might be explained by the fact that the zebrafish genome is smaller compared to human and mouse (Howe et al., 2013) (1/2 the size of human and mouse genomes) but contains a higher number of GRBs distributed across the genome (35% more than human; 30% more than mouse). Long non-coding RNAs and GRBs are both implicated to regulate the expression of genes involved in early development and differentiation (Akalin et al., 2009; Batista and Chang, 2013). Often such genes have a complex expression pattern governed by multiple factors such as enhancers, transcription factors and other non-coding RNAs. A good example is of the Insulin-like growth factor II (*Igf2*) gene in mouse which produces a growth promoting hormone during early gestation (Shen et al., 1988). The *Igf2* is reciprocally imprinted with a long non-coding RNA *H19* which is implicated in cell proliferation and organism growth (Venkatraman et al., 2013). A recent report describes a complex long range interaction between the promoters of *H19* and a novel lncRNA, the Non coding transcript 1 (*Nctc1*) with a shared pool of enhancers to influence the imprinting of the Insulin-like growth factor II (Igf2) in mouse (Eun et al., 2013).

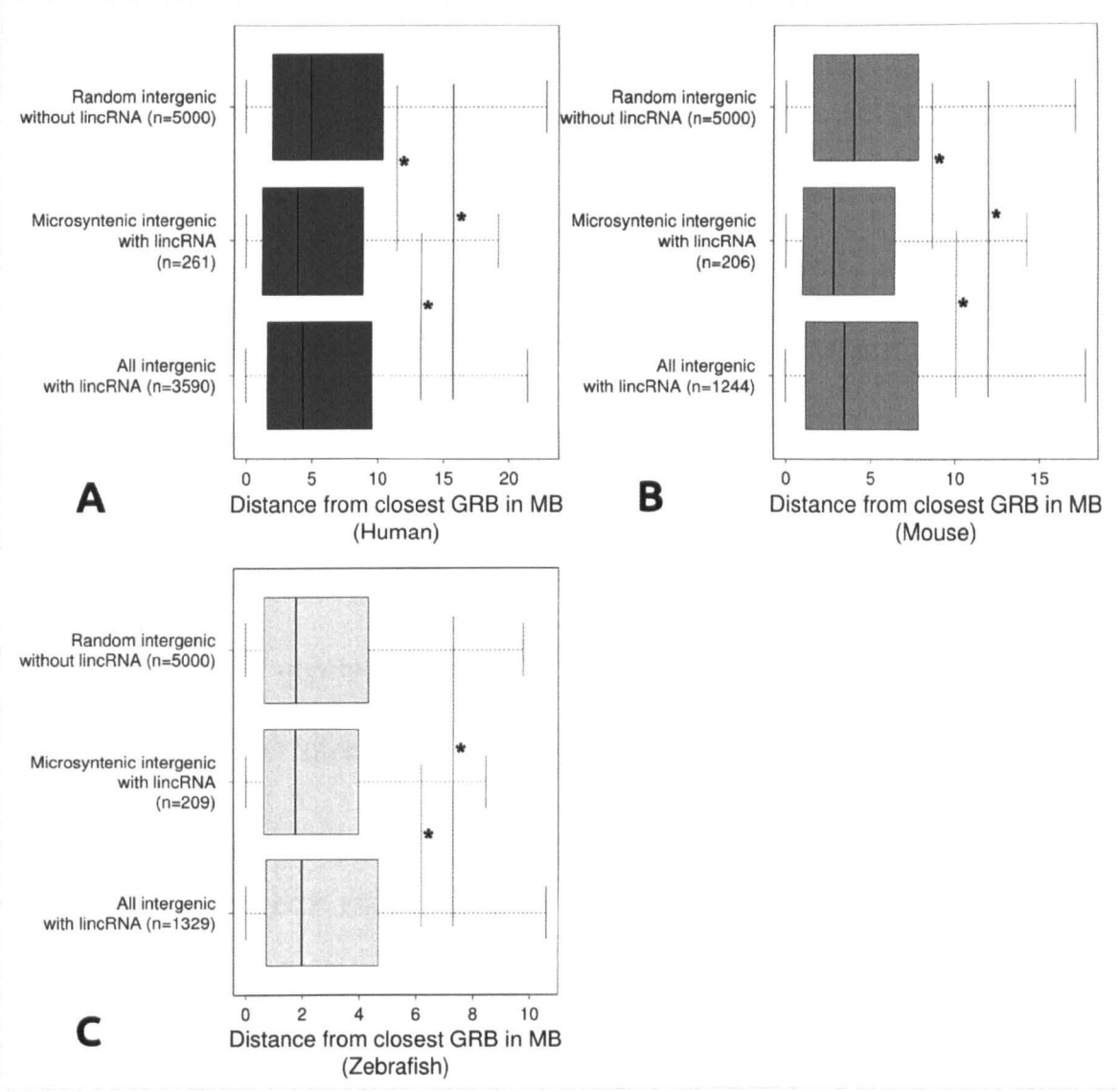


Figure 4.6 Distribution of distance from the closest GRB for intergenic regions containing lincRNAs in **A)** Human **B)** Mouse **C)** Zebrafish. The x-axis represents the distance from the closest GRB and the y-axis represents the different pairs of genomic features.

While experimental evidence is required to assess such associations between coding and non-coding loci, a subset of VMLIGs overlap GRBs which suggests a select set of lincRNAs are transcribed near coding genes whose expression is regulated by elements of a GRB leading to two potential outcomes. Either the GRBs exert their function through transcription of the overlapping lincRNAs or the

lincRNAs and the GRBs have 2 independent, possibly complementary activities. Unfortunately, with the analyses performed until now I have not been able to find a final answer. There may exist complex unknown mechanisms of VML functioning but an intelligible hypothesis could be their involvement in the cis-regulation of flanking coding genes. To test this hypothesis I utilised four computational measures to define cis-regulatory constraints for VMLs

- Expression correlation with flanking coding genes
- Conservation of sequence in the VML/flanking coding interval
- Frequency of regulatory elements proximal to VMLs
- Chromosomal interactions between VMLs and flanking coding genes

To test for the aspects mentioned above the lincRNAs in human, mouse and zebrafish were divided into three categories i) All lincRNAs ii) Vertebrate microsyntenic lincRNAs (VMLs) iii) Vertebrate microsyntenic lincRNAs with retained orientation (VMLRs) with respect to the flanking coding gene. In each dataset the closest coding gene (CCG) was chosen from the two immediate flanking genes of the lincRNA. The choice was defined only by distance for the first dataset (All lincRNAs). For the second dataset (VMLs) the closest orthologous coding gene was considered and for the third dataset (VMLRs) the closest orthologous gene with retained orientation was chosen.

4.3.3 Expression correlation of lincRNAs with their flanking coding genes

Correlation of expression between a lincRNA and its closest coding gene (CCG)

may occur due to sharing of a common regulatory element or by cis-regulation of one feature over another. I compared the expression of lincRNAs and CCGs across multiple tissues in human and mouse. In terms of percentages, I did not find an enrichment for pairs of VMLs/VMLRs and their CCGs to be co-expressed across multiple tissues in comparison to all lincRNAs and their CCGs. However, comparing the expression between putative homologous lincRNA pairs in human and mouse across nine similar tissues, resulted in a slightly higher percentage of VMLR pairs showing correlation as compared to random lincRNA pairs (6% vs 3.5%). Although the difference between these percentages is almost 2 folds, it is not statistically significant and cannot explain the majority of homologous pairs. The lack of expression data for different tissues in zebrafish led me to compare the expression pattern of lincRNAs and CCGs during early developmental stages in whole embryo, for which data are available. Again VMLs, VMLRs and their CCGs do not show any preference to be co-expressed across eight early developmental stages of zebrafish, using the RNAseq data taken from a previously published study (Pauli et al., 2011a). In contrast, the VMLs showed a significant enrichment for positive or negative correlation of expression with their CCGs when I used quantitative data taken from transcriptional start sites defined by CAGE technology (Kodzius et al., 2006) across 12 early developmental stages in zebrafish (Nepal et al., 2013) (**Figure 4.7**). I compared the expression correlation between five set of genomic features which lie adjacent to each other on the zebrafish genome (Spearman correlation score >= 0.9, p-value <= 0.05).

The feature pairs considered are:

- Coding gene with proximal coding.
- Coding gene with proximal coding and retained microsynteny in human, mouse and zebrafish.
- LincRNA with proximal coding gene.
- LincRNA with proximal coding gene, predicted to be microsyntenic in human, mouse and zebrafish (VMLs).
- LincRNA with proximal coding gene, predicted to be microsyntenic with retained orientation in human, mouse and zebrafish (VMLRs).

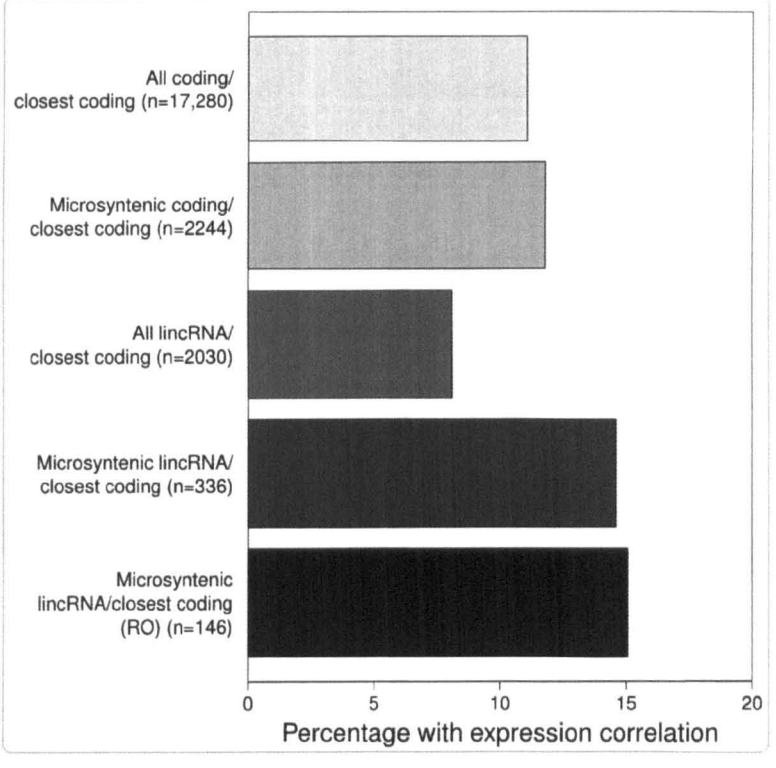


Figure 4.7 Percentage of genomic features (lincRNA/coding or coding/coding) showing expression correlation across twelve developmental stages of zebrafish (spearman correlation score >= 0.9, p-value <= 0.05). RO stands for vertebrate microsyntenic lincRNAs with retained orientation with respect to their closest flanking orthologous coding gene. The x-axis represents the percentage of correlated pairs and the y-axis represents the different pairs of genomic features.

The enrichment is significant for VMLs when compared against all lincRNAs (two sample proportion test: p-value < 0.001) but not when they are compared against microsyntenic coding/coding pairs or VMLRs. These results suggest that, similarly to protein coding genes, microsynteny information helps in selecting co-regulated pairs of coding/noncoding transcripts. A previous study has shown that pairs of coding genes which remain linked with each other (co-linear to each other without interruption by another coding gene) across long evolutionary distances in multiple taxa tend to be co-expressed (Irimia et al., 2012). The pattern of co-expression is suggested to result from mutual interaction between the genes or expression of both the genes being governed by shared regulatory features, common transcription factors or a bidirectional promoter. The positional confirmation of coding and long non-coding pairs across large evolutionary distances can be reasoned on the basis of similar grounds. It is important to note that the enrichment for expression correlation between VMLs, VMLRs and their CCGs could be detected only by CAGE transcription start site abundance data. CAGE technology quantifies sequence tags representing 5' end of RNA molecules (core promoter region) to identify transcription initiation regions genome-wide (Balwierz et al., 2009; Kodzius et al., 2006). A major difference of CAGE from an RNAseq experiment is its ability to quantify the regulation of transcription initiation event. In fact the core promoter region is reported to provide an additional site for regulation of gene expression during development (FANTOM Consortium et al., 2009; Nepal et al., 2013). Further the additional developmental stages present in the CAGE dataset provides better sensitivity to measure variation

of gene expression across smaller time periods. The results show that a subset of VMLs and their CCGs are enriched to be expressed on a similar temporal scale. This enrichment is slightly more pronounced in case of VMLRs but the difference is not statistically significant in comparison to VMLs. This can be interpreted in two ways, firstly that the VMLs influence the expression of their CCG by a direct physical interaction or secondly both the VML and the CCG are share a common regulatory pathway during the early development in the zebrafish. However, I consider the expression enrichment to be a preliminary evidence which needs to be further resolved by experimental validation of candidate the VMLs and their CCGs.

4.3.4 Conservation of sequence in the VML/flanking coding interval

Long non-coding RNAs appear to be more plastic and amenable to evolutionary change in comparison with protein coding genes (Kutter et al., 2012). This is primarily due to a lack of selective pressure on them to retain amino acid codons. Yet for adjacent long non-coding and coding gene pairs retained across long evolutionary distances the sequence conservation in their intergenic interval may reflect a possible functional constraint for the pair to be linked together. This might be due to the presence of conserved transcriptional regulatory elements in the intergenic space shared by both transcriptional units. I used the genome wide phastCons conservation scores to check for sequence conservation in the intergenic intervals between various genomic features. The phastCons program (Pollard et al., 2010) provides base wise conservation scores in the genome of a species by using an alignment of multiple genomes. The mean phastCons scores for the

intergenic space between VMLRs and their CCGs is higher in all the three species in comparison to the intergenic space between conserved coding/coding pairs (**Figure 4.8**). In addition, the VMLRs show a higher conservation in the intergenic space in comparison with all VMLs in human and mouse while, in zebrafish the level of conservation remains similar for the two datasets. Summarizing, the intergenic regions separating VMLRs from their CCGs in all the three species show an evolutionary constraint in terms of sequence conservation. This evolutionary constraint reflects a selection against insertions or deletions which might lead to a possible ablation of shared functional regulatory mechanisms between the VMLRs and their CCGs.

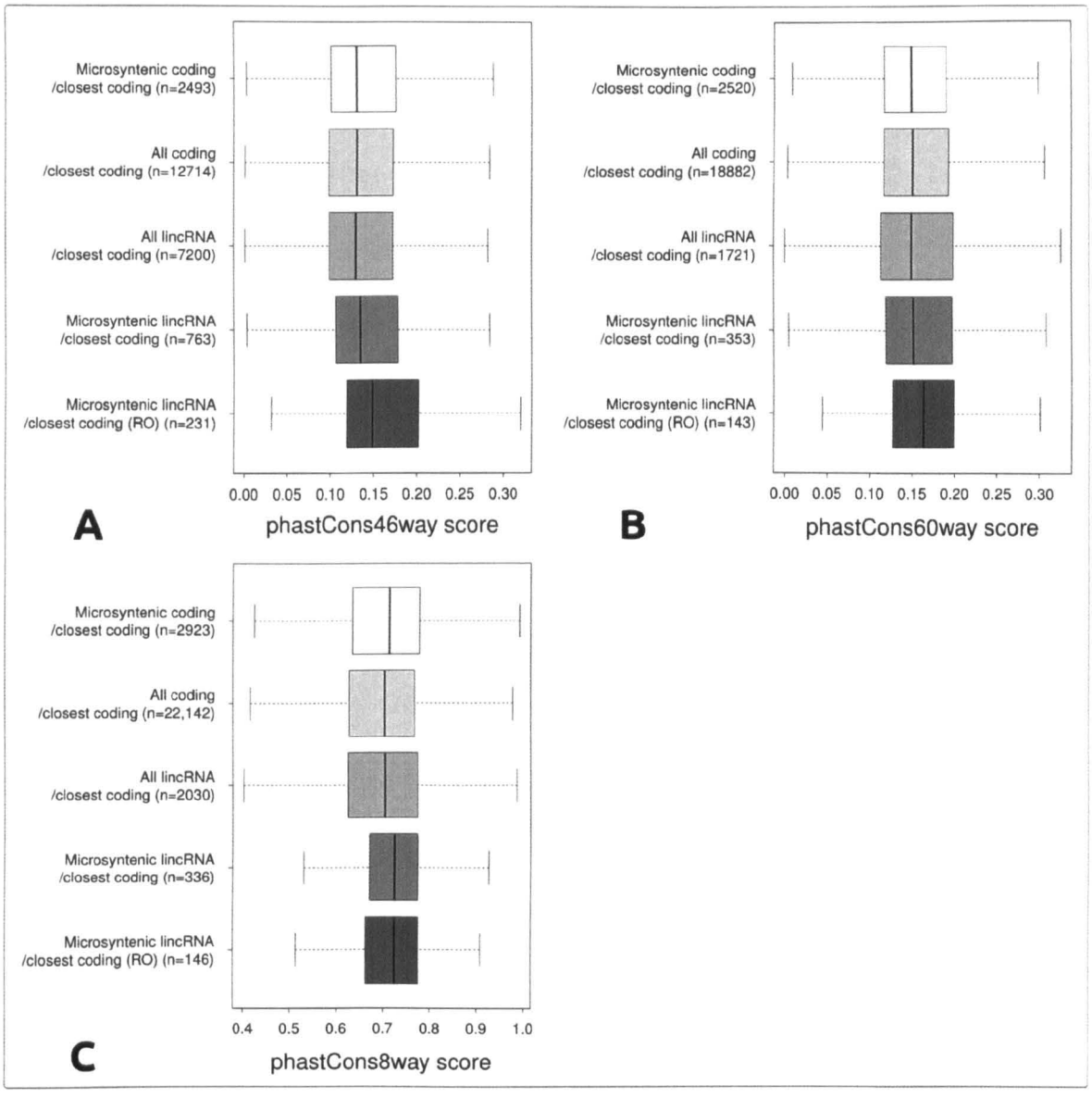


Figure 4.8. PhastCons conservation scores for intergenic intervals between different genomic features (lincRNA/coding or coding/coding) in **A)** Human **B)** Mouse **C)** Zebrafish. The x-axis represents the phastCons conservation scores and the y-axis represents the different pairs of genomic features.

4.3.5 Frequency of regulatory elements proximal to VMLs

The vertebrate genome can be described as a transcriptional mosaic where regulatory features like enhancers and insulators are closely knit with their target genes guiding their tissue and stage specific expression. Regulatory features like

enhancers and insulators may aid in the retention of a local gene order by controlling the expression of their neighboring genes, guiding the development and differentiation of tissues and organ systems (Kolovos et al., 2012). I specifically looked for the frequency of enhancers and insulators near lincRNAs to understand if they show an enrichment near VMLs or VMLRs. There are previous reports which show the enrichment of regulatory elements in large genomic regions, deficient in coding genes (Bagheri-Fam et al., 2001; Montavon et al., 2011). Hence before looking for the enrichment of regulatory elements I checked for the length distribution of the intergenic intervals separating lincRNAs from their CCGs (**Figure 4.9**). The lincRNAs are separated by longer intergenic intervals from their CCGs in comparison to coding genes (Students t test, p-value < 2e-16 for all species). Within lincRNAs, the VMLs and VMLRs are separated by even longer intervals from their CCGs in comparison to all lincRNAs. A larger intergenic interval between lincRNAs and their CCGs argues for the higher probability of a regulatory feature to occur in the interval by chance.

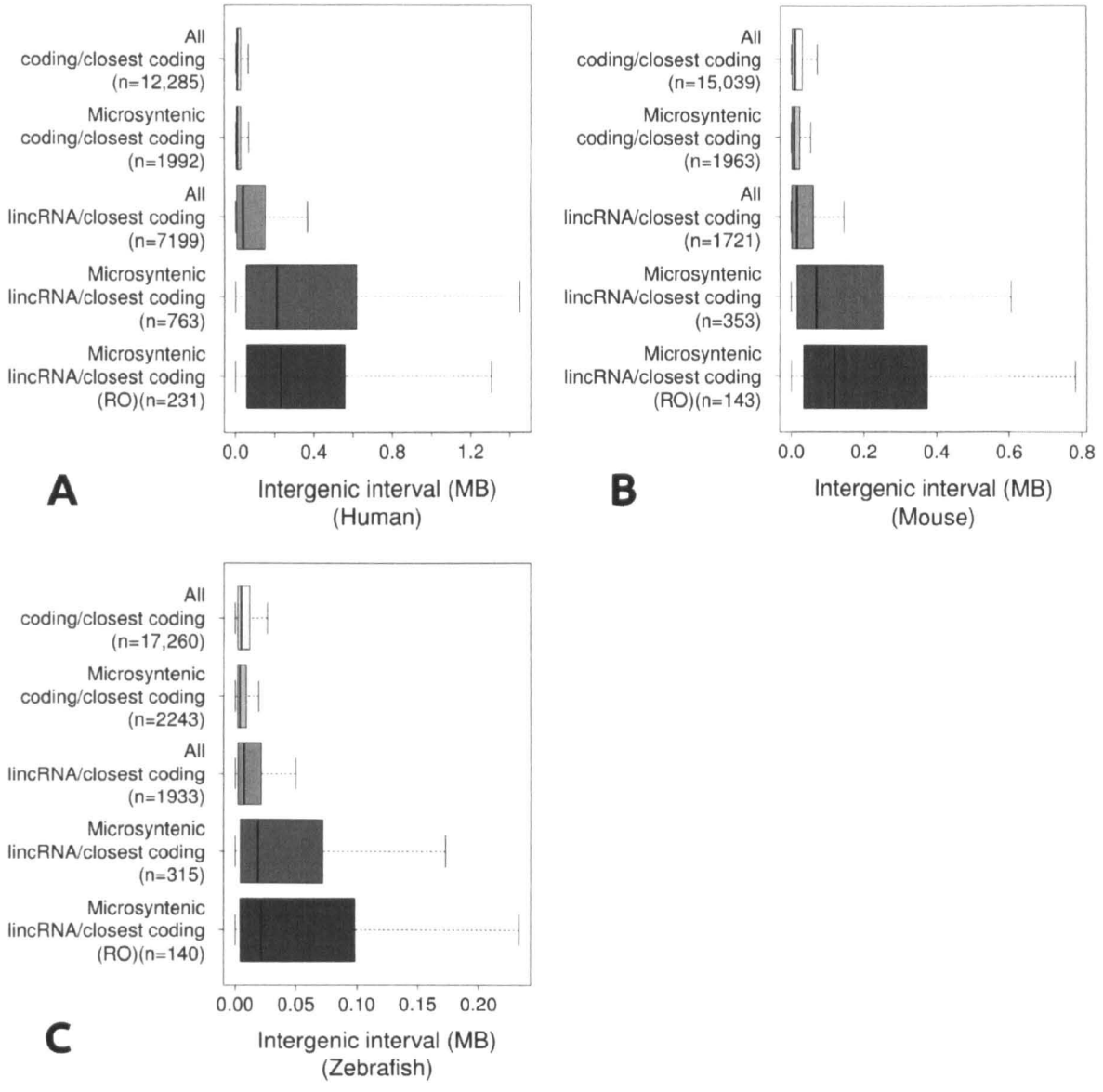
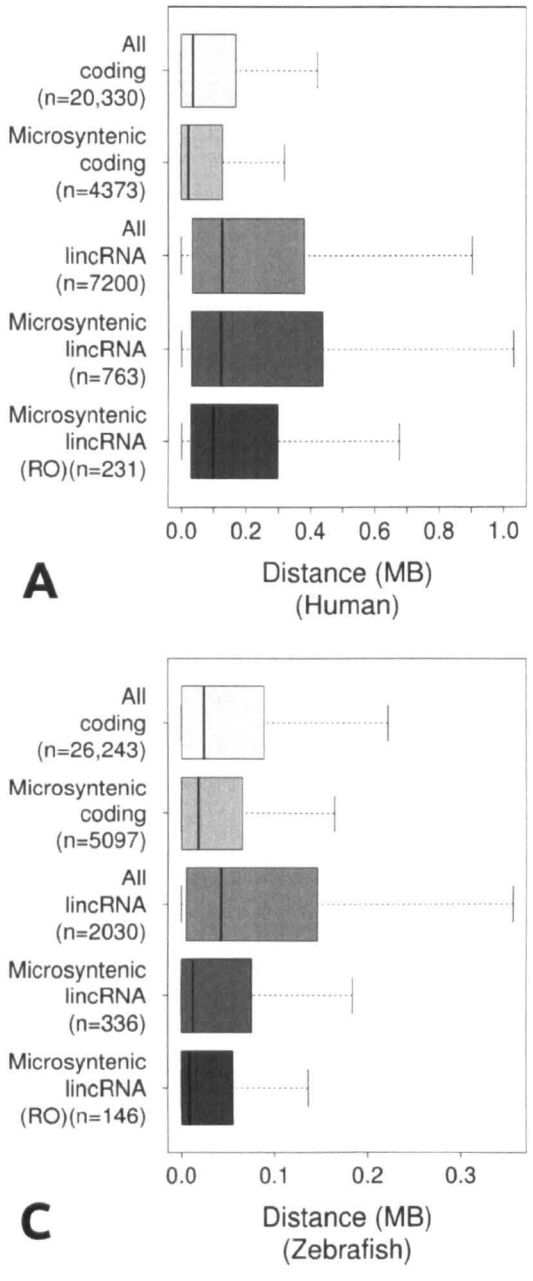


Figure 4.9. Distribution of the intergenic interval length between different genomic features in **A)** Human **B)** Mouse **C)** Zebrafish. The x-axis represents the size of intergenic intervals and the y-axis represents the different pairs of genomic features.

Long non-coding RNAs are known to be associated with enhancer elements with two recent reports providing a strong experimental evidence of lincRNAs acting as enhancers (Li et al., 2013a; Yang et al., 2013a). In the past a large scale screening associated numerous transcribed lncRNA regions with enhancer activity (Ørom et al., 2010a). In fact a few studies have highlighted a complex interplay between

enhancers and long non-coding RNAs to assert a transcriptional control over adjacent coding genes (Berghoff et al., 2013; Eun et al., 2013; Korostowski et al., 2011). To check for the proximity of enhancer elements with respect to VMLs and VMLRs I calculated the distance of each lincRNA from its closest mapped conserved active enhancer (see Material and methods and **Figure 4.10**). Except for zebrafish, the VMLs do not show an enrichment to lie near enhancer elements in comparison with all coding and microsyntenic coding genes. While the results for zebrafish suggest an association between enhancer elements and VMLs the same conclusion cannot be drawn for human and mouse. I conclude that the enrichment of VMLs to lie near enhancer elements cannot be observed in all the three species of interest. Thus the hypothesis of enhancer elements influencing the retention of lincRNA microsynteny remains inconclusive.



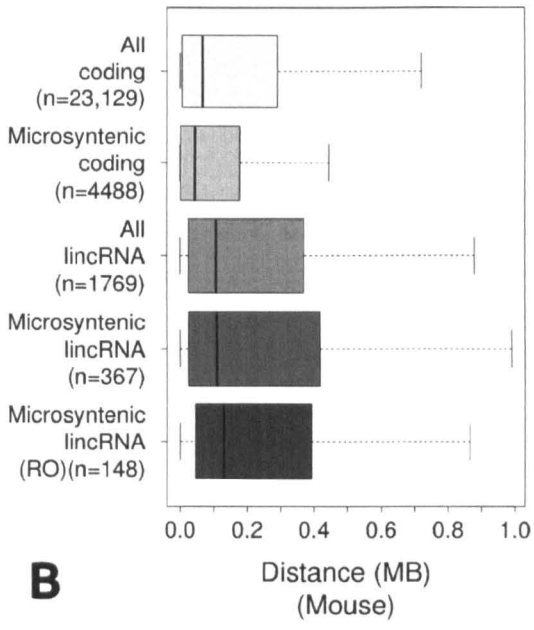


Figure 4.10 The distance of closest conserved active enhancer mark from lincRNAs and coding genes in **A)** Human **B)** Mouse **C)** Zebrafish. The x-axis represents the distance from the closest enhancer mark and the y-axis represents the different genomic features.

In vertebrates the word insulator has often been associated with the binding of CCCTC-Binding Factor (*CTCF*) protein on the genome which appears to be counterintuitive. Classically the *CTCF* protein is reported to bind *cohesin* and

mediate transcriptional regulation by insulating promoters from distal enhancers (Wendt et al., 2008). Recent evidences associate the *CTCF* protein with another mechanism, that is guiding long range chromatin interactions by formation of chromatin loops. A well known example of such a mechanism is the *CTCF* mediated interactions bringing regulatory sequences of the *Insulin* (*INS*) gene close to the Synaptotagmin VII (*SYT8*) gene to coordinate the expression of both the genes for regulation of insulin secretion (Xu et al., 2011). In fact formation of *CTCF* mediated chromatin loops is observed genome wide in mouse embryonic cells (Handoko et al., 2011). To inquire about a possible role of *CTCF* mediated interaction between a lincRNA and its CCG, I counted the number of *CTCF* binding sites lying in the intergenic interval of different genomic features (**Figure 4.11**). Interestingly VML and VMLR intervals show an enrichment for *CTCF* binding sites when compared to all lincRNAs, coding genes and microsyntenic coding genes (Mann-whitney test p-value < 0.001). The results show that a significant percentage of VMLs and VMLRs have at least one *CTCF* binding site per megabase of intergenic interval separating them from their CCG (VML: Human 48%, Mouse 56%; VMLR: Human 56%, Mouse 56%) in comparison to all linRNAs (Human: 30%, Mouse: 33%), all coding genes (Human: 21%, Mouse: 31%) and microsyntenic coding genes (Human: 19%, Mouse: 33%). This points towards a possible implication of *CTCF* and its associated proteins mediating long range interactions between VMLs and their proximal coding genes.

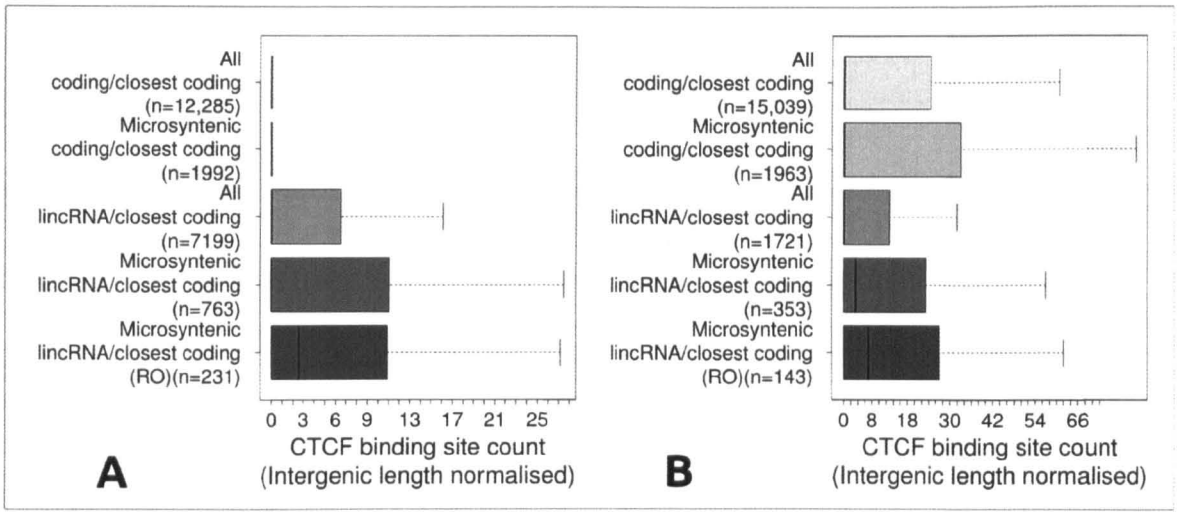


Figure 4.11 The distribution of *CTCF* binding sites in intergenic intervals between different genomic features (long non-coding/coding, coding/coding) in **A)** Human **B)** Mouse. The x-axis represents the distance from the closest CTCF binding site and the y-axis represents the different pairs of genomic features.

4.3.6 Chromatin interactions between lincRNAs and proximal coding genes

While expression correlation, sequence conservation and regulatory feature proximity, to a certain extent, provide support to the hypothesis of cis-regulatory constraint in VMLs and their flanking coding genes, a direct evidence of lincRNA and coding gene interaction is still lacking. Imprinting and tethering are two known functions of lncRNAs, which are related to lncRNAs regulating the expression of proximal coding genes (Gabory et al., 2010; Jeon and Lee, 2011). However such mechanisms may require chromosomal looping to bring a lncRNA in physical proximity to its target gene. Based on chromatin interactions a recent report showed the presence of several cis-interacting regions flanking the *Sox9* gene which overlap lincRNA loci, thus suggesting a role of the lincRNAs in mediating the regulation of *Sox9* gene (Smyk et al., 2013). Another example is of

the *HOTTIP* lincRNA which lies upstream of the *Hoxa* cluster and is dependent on chromosomal looping to achieve physical proximity followed by activation of its target *Hoxa* cluster genes (Wang et al., 2011). Hence I decided to compare the genomic coordinates of interacting chromosomal locations (mapped by cross-linking experiments) with those of lincRNAs and their proximal coding genes. To perform the analysis I used Hi-C and ChIA-PET data from human and mouse embryonic stem cells. Both Hi-C and ChIA-PET are techniques to map long range chromatin interactions within a genome (Fullwood et al., 2009; Lieberman-Aiden et al., 2009). While ChIA-PET failed to identify interactions between the VMLs and their closest coding genes, Hi-C data predicted that 70-90% of all pairs of genomic features (coding/coding and coding/long non-coding in human and mouse) are implicated in interactions. However I could not find any specific enrichment for interaction scores associated to VMLs or VMLRs. Hence I was unable to obtain a direct evidence of interaction between a VML and its CCG. Yet lack of interacting evidence only suggests an absence of direct physical contact between a lincRNA and a coding gene. There are many other mechanisms like binding to transcription factors, altering of the chromatin state and inhibition of splicing, through which a lincRNA can exert its regulatory potential on a neighboring coding gene (Kornienko et al., 2013).

4.3.7 Specific examples of microsyntenic lincRNAs

Manual inspection of the SynLinc pipeline results led me to identify specific lincRNAs reported in prior publications (**Table 4.2**). The interesting part was the

pipeline being able to identify the putative zebrafish orthologs of known mammalian lincRNAs. Furthermore I observed a lack of conservation of splicing pattern or transcript length between the predicted orthologs of the known and characterised lincRNAs. An interesting observation from the manual inspection was the failure of the pipeline to predict the mouse ortholog of the human *Xist* lincRNA. The mouse *Xist* overlaps a neighboring coding gene, hence is filtered by the pipeline, which in the actual version is focused around completely intergenic noncoding transcripts. Currently I am adding additional modules to the SynLinc pipeline which extend the detection of putative syntenic association also to the lncRNAs overlapping coding genes. I expect the additional modules to provide further insights into the conservation of lncRNAs across multiple species.

Symbol	Name	Function	Mechanism	Reference
CRNDE	Colorectal Neoplasia Differentially Expressed	Neuronal differentiation, cancer metastatis	Epigenetic modification	(Ellis et al., 2012)
H19	Human 19	Early growth and development	*Cis* and *Trans* regulation	(Eun et al., 2013; Venkatraman et al., 2013)
HAR1A	Human accelerated region 1A	Early cortical development	Co-localisation with cell signaling protein	(Pollard et al., 2006b)
ANCR	Angelman syndrome chromosome region	Cell proliferation	Suppression of transcriptional regulators	(Kretz et al., 2012)
SNHG15	Small Nucleolar RNA Host Gene 15	Response to chemical stress	Unknown	(Tani and Torimura, 2013)
Tsix	XIST Antisense RNA	X chromosome reactivation	Cis-regulation of *Xist* gene	(Ohhata et al., 2011)

Table 4.2 List of vertebrate microsyntenic lincRNAs predicted by the SynLinc pipeline which are reported in prior published studies.

Discussed below are three examples of VMLRs with known functions in at least one of the analyzed species. First example is represented by the human accelerated region (*HAR1A, HAR1B*) lincRNA (**Figure 4.12**) genes which are derived from genomic regions with conserved sequence in vertebrates but known to evolve rapidly since the divergence of humans from apes (Pollard et al., 2006c). These genes are reported to be expressed in the developing neocortex during human embryonic development and co-localize with *Reelin*, a protein implicated in schizophrenia and aging (Pollard et al., 2006b). Based on the orientation and proximity of the YT521-B homology Domain Family 1 gene (*YTHDF1*) the SynLinc pipeline has been able to predict the putative homologs of human *HAR1A* and *HAR1B* in mouse and zebrafish (**Figure 4.12 B, C**). The *YTHDF1* gene contains a *YTH* RNA-binding domain which is a RNA-binding domain involved in splicing of vertebrate genes (Zhang et al., 2010). In mouse the putative *HAR1* locus contains five annotated lincRNA transcripts while in zebrafish there are only two annotated lincRNA transcripts proximal to the *YTHDF1* gene. An additional support for the retention of microsynteny comes from the presence of *mir-124* family genes between the *HAR* lincRNA homologs and *YTHDF1* in all the analysed species. In a previous analysis I have detected two zebrafish lincRNAs overlapping the *mir-124a5* gene and a mouse lincRNAs overlapping the *mir-124a1* gene hence it is interesting to note this close association of this miRNA gene with lncRNA transcripts.

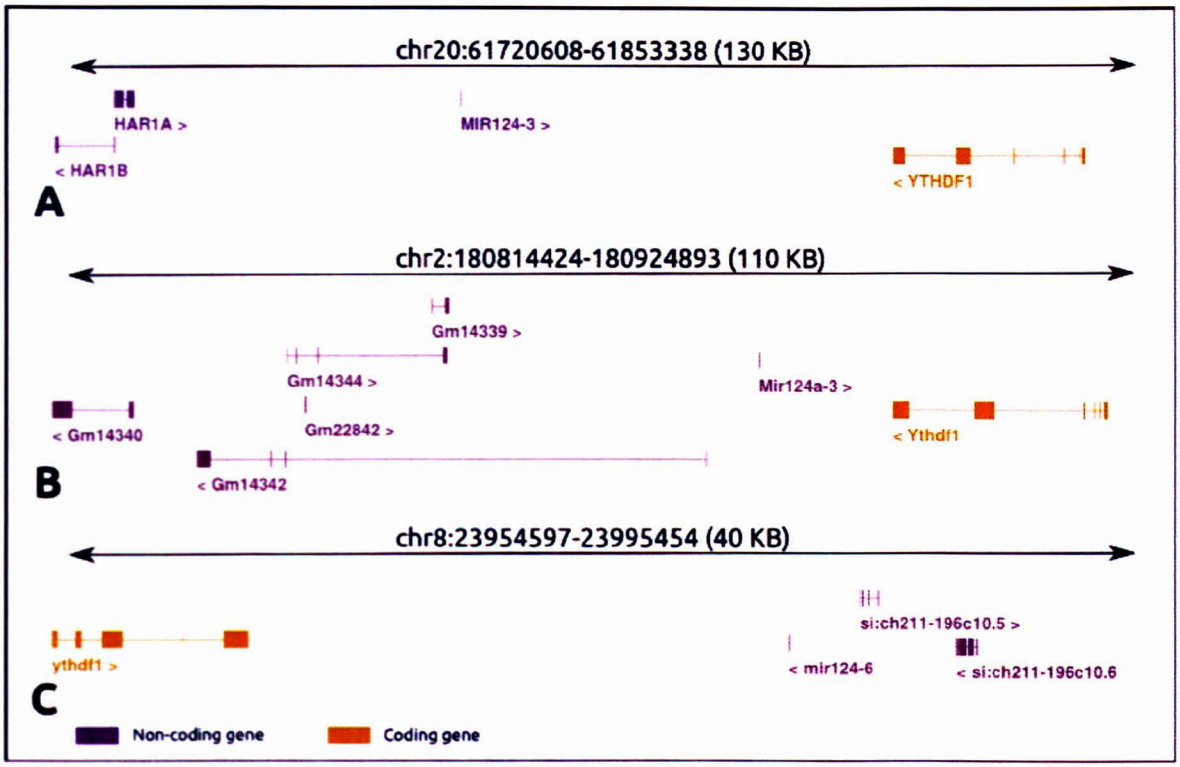


Figure 4.12 Putative *linc-HAR1A*, *linc-HAR1B* orthologues predicted by the SynLinc pipeline in **A)** Human **B)** Mouse **C)** Zebrafish.

The second example is that of the lincRNA Angelman syndrome chromosome region (*ANCR/DANCR)* (**Figure 4.13**) which maintains the undifferentiated state of human epidermal progenitor cells (Kretz et al., 2012). The *DANCR* lincRNA lies close to the Ubiquitin specific peptidase 46 (*Usp46*) gene which is a deubiquitination enzyme reported to act as a tumour suppressor by up-regulating the PH domain and Leucine rich repeat Protein Phosphatases gene (*PHLPP1*) in human colon cancer cell lines (Li et al., 2013b). Additionally, this gene is also implicated in the regulation of glutamate receptor expression the ventral nerve cord in *C. elegans* thus modulating its synaptic strength (Kowalski et al., 2011). The *DANCR* orthologs in all the three species overlap a snoRNA (*snoRNA26*) which is

an additional indicator of conservation. This lincRNA is reported to regulate the differentiation of human mesenchymal stem cells into osteoblasts by binding with Enhancer Of Zeste Homolog 2 (*EZH2*) protein and inhibiting the expression of Roughex 2 gene (*Rux2*) (Zhu and Xu, 2013). While the predicted mouse lincRNA is already named *DANCR* by the Ensembl gene classification pipeline, the zebrafish homolog remains unannotated.

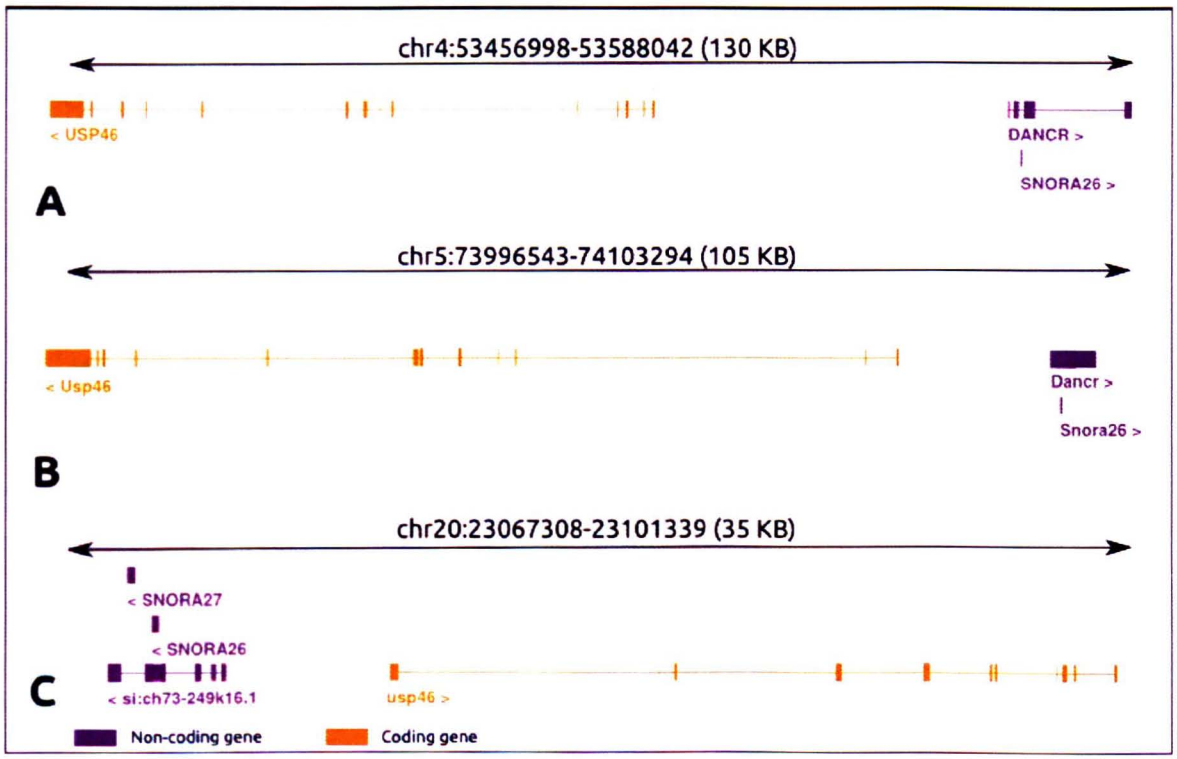


Figure 4.13 Putative *linc-DANCR* orthologues predicted by the SynLinc pipeline in **A)** Human **B)** Mouse **C)** Zebrafish.

The final example is of the recently reported Small Nucleolar RNA Host Gene 15 (*SNHG15*) lincRNA (**Figure 4.14**) which is expressed, with a very short half life, in response to chemical agents in human cell lines (Tani and Torimura, 2013). This

lincRNA was predicted to be microsyntenic based on the homology of its downstream coding gene *Myo1G* which is reported to regulate the elasticity of haematopoietic cells (Olety et al., 2010). Except for a few well studied candidate lncRNAs like the *XIST*, *HOTTIP*, *AIR*, *H19* and *Kncq1ot1* there is no conclusive experimental evidence for wide spread cis-regulatory mechanism of lncRNAs. The predicted VMLs and VMLRs are the first set of lincRNA candidates predicted to show conserved association with a coding gene in human, mouse and zebrafish. These results demonstrate the utility of the SynLinc pipeline to employ microsynteny as a paradigm and reduce the lincRNA search space in organisms with sequenced genomes.

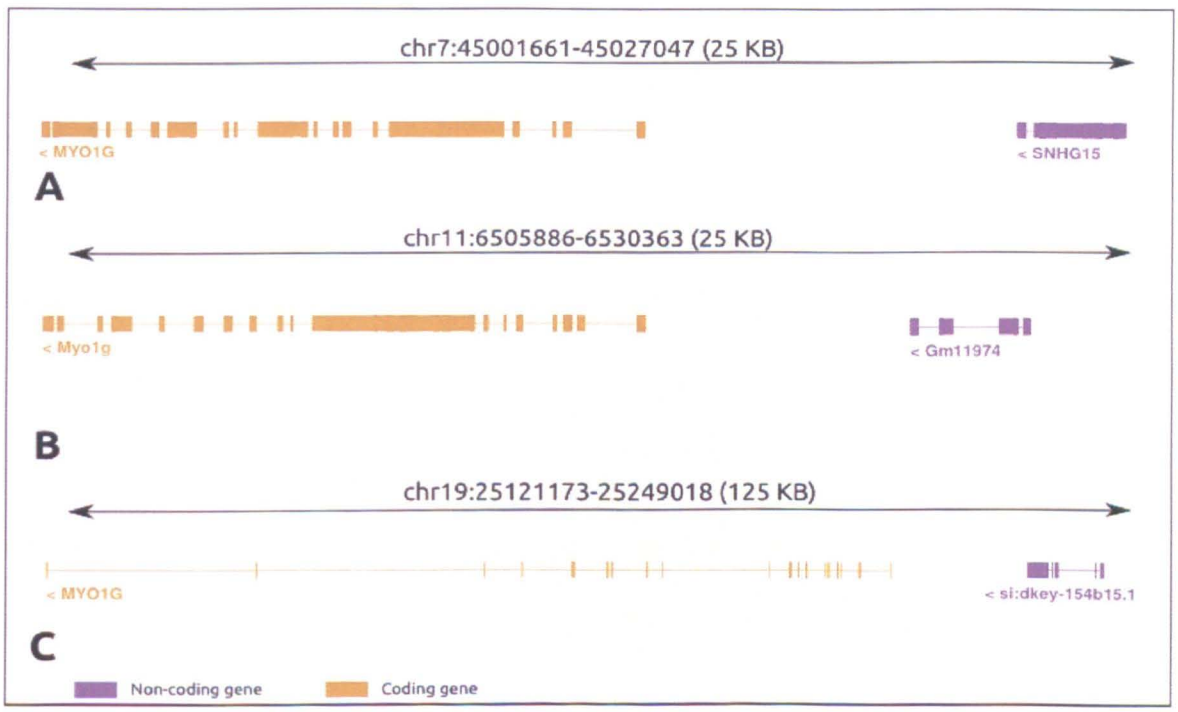


Figure 4.14 Putative *linc-SNHG15* orthologues predicted by the SynLinc pipeline in **A)** Human **B)** Mouse **C)** Zebrafish.

4.4. Conclusion

The current scenario for computational prediction of lincRNA mechanism and function lies between what we already know about coding genes and how well we can extrapolate that knowledge to identify lincRNAs and predict their function. Such an approach is focused towards proving *"what the lncRNAs are not"* rather than understanding *"what they are"*. While analyses of sequence, secondary structure and mRNA expression pattern have predicted conserved lncRNAs, weighing upon their evolutionary mobility most lncRNAs are expected to morph beyond recognition between phylogeneticaly distant species. The use of microsynteny removes the bias against confinement of a lncRNA to norms of sequence, size or structure and allows for position as the only criteria to predict homology.

To identify putative microsyntenic lincRNAs between two species I developed a computational pipeline named SynLinc. The pipeline is designed to compare the lincRNA population in two genomes in the context of position and orientation of coding genes present in the genomic neighborhood at a defined distance threshold. I used the pipeline to identify putative vertebrate microsyntenic lincRNAs (VMLs) in human, mouse and zebrafish. The predicted VMLs are not conserved due to a random placement on the genome and a small subset show correlation of expression in comparison to their proximal coding genes during early development in zebrafish. The VMLs which retain their orientation with respect to a flanking coding gene (VMLRs) show a higher order of sequence

conservation in the intergenic interval between the coding and non-coding sequence.

Examples of a few published lincRNAs predicted as VMLRs show the capability of the pipeline to identify unreported orthologs of mammalian lincRNAs in zebrafish making it a useful resource to detect lincRNA conservation in different datasets. While such an approach may not be ideal in many cases, it is advantageous to reduce the search space and then focus in-depth on a smaller predicted microsyntenic subset. The SynLinc pipeline is able to perform this task in a quick and efficient manner requiring a minimal level of user input. Manual evaluation of a few examples of predicted microsyntenic lincRNAs showed a lack of conservation of splicing pattern and length even in those cases where the lincRNAs remain linked in the same orientation with a coding gene. The pipeline is capable of predicting higher numbers of putative microsyntenic lincRNAs if run with a more relaxed distance threshold. Without experimental validations a strong conclusion cannot be drawn on the predictions, yet the current evidence suggests the microsynteny approach to be well suited to identify lincRNA candidates which may be under the influence of co-regulatory mechanism with respect to their proximal coding genes.

Chapter 5

Identification of long non-coding RNAs in pancreatic islet cells of zebrafish

5.1 Introduction

5.1.1 Zebrafish as a model system to study human diseases

The sequencing of the human genome led to the identification of mutations and polymorphisms in numerous genomic loci, implicated in various mendelian disorders (Begum et al., 2012; Costa et al., 2013). An underlying challenge is to unravel the molecular mechanism relating a genomic locus to pathophysiology. Precise experimental evidences from comparative genomics (Wallace et al., 2007; Zheng-Bradley et al., 2010) have given support to the premise of evolutionary conserved pathogenesis mechanisms, encouraging the use of animal models to study human diseases. The mouse model has caught the maximum attention of the scientific community, with examples of a disease mechanism being defined in mouse before humans (Kljuic et al., 2003) and similar phenotypes generated for loss-of-function mutations in genes orthologous to human (Al-Hasani et al., 2005). Amongst other organisms particularly *Mus musculus* (mouse), *Danio rerio* (zebrafish), *Drosophila melanogaster* (fruitfly) and *Caenorhabditis elegans* (worm) have

gained wide acceptance as model organisms. The use of zebrafish as a model organism to study molecular mechanisms or a disease state has risen to prominence recently because of the following reasons.

- Zebrafish is a vertebrate with a sequenced genome and ~70% of human protein coding genes have a known ortholog in zebrafish (Howe et al., 2013).
- Several organ systems in zebrafish are notably similar to human (Goldsmith and Jobin, 2012).
- Zebrafish embryos have a translucent body which aids in monitoring its organ systems and cellular development *in vivo*.
- Zebrafish are highly fecund and a pair of zebrafish can produce up to 200 embryos per clutch.
- Large scale forward-genetics approaches can be carried out in the organism (Driever et al., 1996; Haffter et al., 1996).
- Large scale reverse genetics approaches like TILLING, retroviral mediated mutagenesis, zinc finger nucleases, TALENs and CRISPR and morpholino knockdown can be efficiently carried out in the zebrafish system (Bedell et al., 2011; Blackburn et al., 2013; Doyon et al., 2008; Kettleborough et al., 2013; Petzold et al., 2009; Wienholds et al., 2003).
- Finally transgenesis experiments can be easily performed in zebrafish, thus allowing it to effectively function as a model for various human diseases (Becker and Rinkwitz, 2012; Liu and Leach, 2011).

The zebrafish has a well developed gastrointestinal system which is homologous to

mammals (Wallace et al., 2005). This stark similarity has led to several diseases of the gut being modeled in zebrafish like liver cancer (Lam and Gong, 2006), pancreatic cancer (Park et al., 2008) and inflammatory bowel disease (Brugman et al., 2009). Further, the high fecundity, transparency, and ease of imaging of the zebrafish embryos has encouraged genetic screening approaches to identify fish phenotypes for inheritable pathologies like polycystic kidney disease (Sun et al., 2004) and dilated cardiomyopathy (Xu et al., 2002).

5.1.2 Zebrafish as a model system to study the molecular mechanisms of type 2 Diabetes

The similarity in organ systems and the potential for large scale genetic screening makes zebrafish an ideal model organism to understand the complexities of human diseases associated with heredity and lifestyle such as diabetes (Seth et al., 2013). The Type 2 *diabetes mellitus* (T2DM) is a complex metabolic disorder associated with a high glucose level in the body caused by resistance to cellular uptake of insulin, and deficiency in insulin production (Taylor, 1999). It is predicted to be an emerging epidemic amongst the elderly (Kesavadev et al., 2003; Zeyfang and Bahrmann, 2013) and prevalent, also amongst the youth and children (Van Name and Santoro, 2013; Pettitt et al., 2013). In fact the World Health Organisation (WHO) estimates around 60 million people in the European economic region affected with the disease and the rate of deaths by T2DM to be doubled by the year 2030 (http://www.euro.who.int/). The primary cause of the disease is the resistance of body cells against the insulin hormone resulting in lack

of glucose regulation and dysfunction of the insulin secreting β-cells in the pancreatic islets (Kahn et al., 2006). The whole mechanism of glucose regulation and insulin secretion is a complex process which involves the interplay of proteins associated with multiple disease related pathways like Alzheimer (Dash, 2013), obesity (Kahn et al., 2006) and atherosclerosis (Stöhr and Federici, 2013). There is also a direct involvement of circulating hormones and nutrients (Braun et al., 2012) as well small non-coding RNAs like the miRNAs (McClelland and Kantharidis, 2014). The zebrafish pancreatic islet cell organisation is similar to humans, comprising of β-cells (secreting insulin) surrounded by α-cells (secreting glucagon), δ-cells (secreting somatostatin) and ε-cells (secreting ghrelin) (Kim et al., 2006). The ability of zebrafish to regenerate chemically or surgically removed pancreas makes it an ideal choice to study the proliferation of pancreatic cells especially the regeneration of β-cells (Moss et al., 2009). Recently reverse genetic studies have identified genes important in zebrafish pancreatic development whose mammalian orthologs have similar functions. Amongst them are *ISL LIM* homeobox 1 (*Isl1*) and *ISL LIM* homeobox 2 (*Isl2*) genes, which are involved in formation of pancreatic cells (Wilfinger et al., 2013) and Adehyde dehydrogenase 1 *(ALDH1)* gene inducing endocrine differentiation in the pancreas (Matsuda et al., 2013). There is also a report of specific cis-regulatory elements in zebrafish coordinating the distinct expression pattern of Eukaryotic Translation Termination Factor 1a (*ETF1a*) gene which guides the expansion of pancreatic progenitor cells (Pashos et al., 2013).

5.1.3 Role of long non-coding RNAs in pancreatic development and the islet-cell transcriptome in zebrafish

Several reports in the past have indicated a role of multiple miRNAs in pancreas development and differentiation thus relating them to the likely prognosis of diabetes (Guay et al., 2012). The down-regulation of the *H19* gene in mice with gestational diabetes mellitus compared to the wild type (Ding et al., 2012), probably caused by an abnormal methylation pattern in the *H19* locus (Shao et al., 2008), demonstrated, for the first time, the association of a long non-coding RNA (lncRNA) with the diseased state. Another study reported around 1100 lncRNA genes expressed in human pancreatic islet cells with possible implication in T2DM (Morán et al., 2012). Majority of the lncRNAs (70%) were shown to have a lncRNA ortholog in mouse also expressed in the islet cells and many lncRNAs mapped to genetic loci underlying diabetes susceptibility. Further most of the lncRNAs are found to lie near protein coding genes which themselves are islet specific. Depletion of a candidate lncRNA in the human β-cells led to the dysregulation of the GLIS family zinc finger (*GLIS3*) which is a key regulator of insulin transcription (ZeRuth et al., 2013) suggesting a regulatory potential of lncRNAs in T2DM metastasis. These reports indicate the need for a better understanding of the role of non-coding RNAs in pancreatic cell development and differentiation, especially lncRNAs. The study of coding genes and lncRNAs in the context of pancreatic development and differentiation in zebrafish may provide novel insights into the various mechanisms influencing the pancreatic metabolism. Hence I have assembled and annotated the islet-cell transcriptome at 72 hours post

fertilisation in zebrafish. I wanted to examine the coding gene repertoire which is differentially up-regulated in the islet cells. Further I also wanted to identify differentially up-regulated lincRNAs in the islet cells of zebrafish and select candidate lincRNAs for further experimental validation. To achieve my objectives I used the Annocript pipeline previously developed by me to annotate the coding and long non-coding transcripts and further employed the SynLinc pipeline to identify potential microsyntenic lincRNAs, expressed in islets of human and zebrafish.

5.2 Materials and Methods

5.2.1 RNA extraction and sequencing

The RNAseq data was generated in the laboratory of my external supervisor Dr. Ferenc Müller. All the experiments for islet cell enrichment and RNA extraction were performed by his doctoral student Irene Miguel-Escalada. The protocol for zebrafish islet enrichment has been taken from a previous article describing isolation of embryonic hearts in zebrafish (Burns and MacRae, 2006). A transgenic line expressing mCherry fluorescent protein in insulin-producing β-cells was used to identify zebrafish islet cells (Pisharath et al., 2007). Homozygous transgenic adult zebrafish males were outcrossed with AB*WT female fish in a 1:2 ratio. The embryos acquired from the cross were fragmented and Intact mCherry+ islets cells were identified under fluorescent light. To reduce the presence of other pancreatic tissues, the islets were accumulated in a clean drop of L-15 Medium (**Figure 5.1**).

From 800 embryos, approximately 200 islets were recovered and pelleted. Total RNA was extracted from whole embryos at 72 hours and islet cells with RN*easy* Micro kit (QIAGEN, UK), following manufacturer's instructions. RNA integrity and yield was evaluated using Agilent RNA Pico 6000 kit. RNA sequencing was done on the Illumina platform (50 base reads, paired-end sequencing, two samples). Quantitative PCR was performed using TaqMan® Gene Expression Assays for insulin and glucagon (islet-specific), exocrine pancreas (trypsin) and liver (lfabp/fabp1a), heart (myl7) and retina and diencephalon (six3b). The islet sample showed a significant fold enrichment for the islet specific genes (406 fold insulin, 374 fold glucagon) in comparison to whole embryo. However there was a enrichment detected for trypsin (76 fold) suggesting some contamination of exocrine pancreatic tissue in the islet sample. Contamination from other tissues was observed to be negligible.

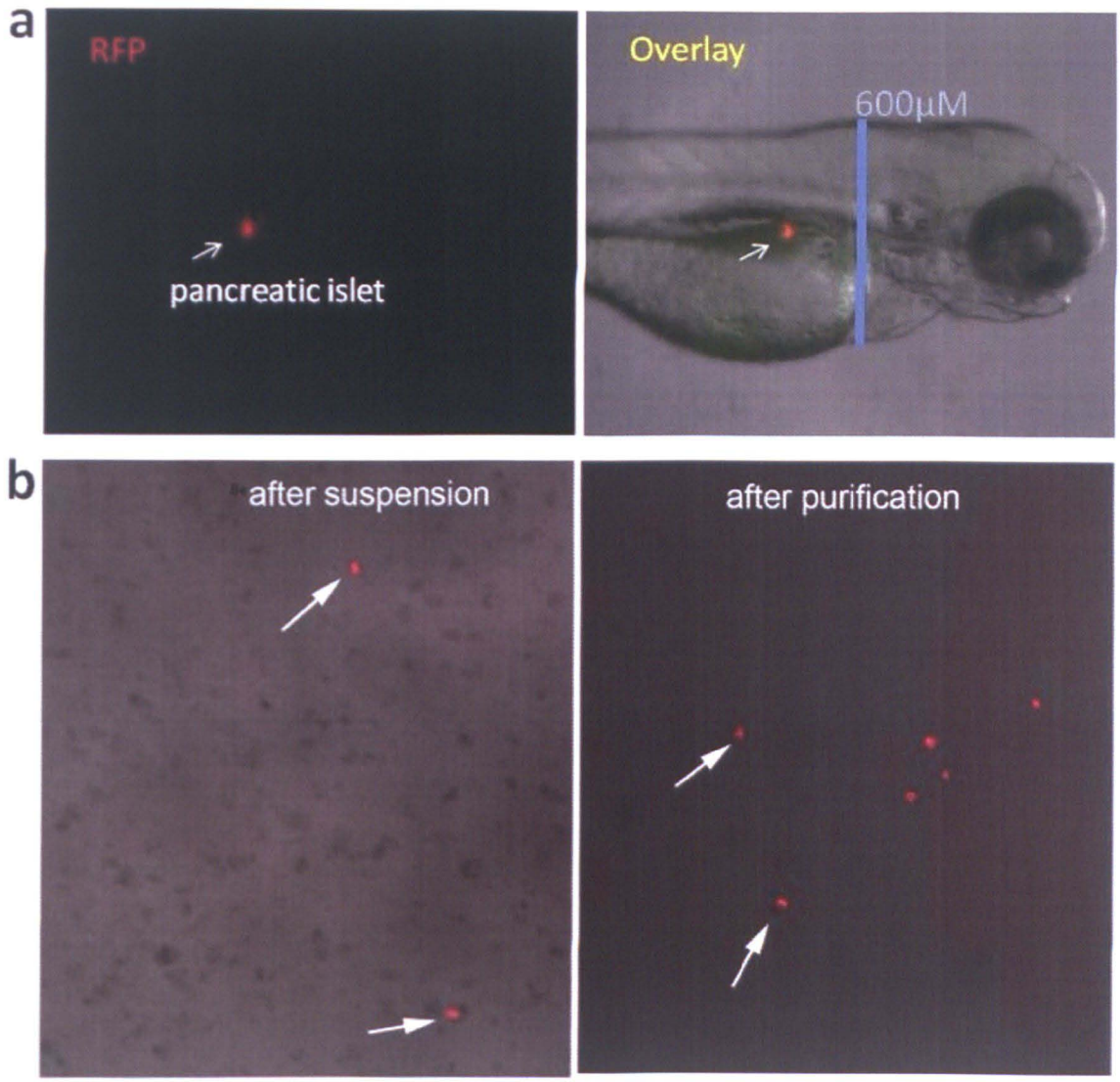


Figure 5.1 Enrichment of pancreatic islets from zebrafish embryos. **A)** Lateral view of a 3 dpf embryo from Tg (ins-mCherry)jh2 line. **B)** Intact zebrafish mCherry + islets (white arrows) from Tg (ins-mCherry)jh2 line in suspension with embryo fragments after mechanical disruption (left panel). Isolated zebrafish islets after collection and transfer into a clean drop of medium with non-pancreatic tissue (right panel). The figure is taken from the doctoral thesis work of Irene Miguel-Escalada.

5.2.2 Quality filtering, mapping and assembly of sequenced reads

The raw sequencing reads from 72 hour post fertilization (72 hpf) whole embryos and islet cells were processed with the Trimmomatic program (Lohse et al., 2012)

to trim low quality bases, filter reads with low quality and filter reads smaller than 36 bases after trimming (parameters: ILLUMINACLIP::2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 HEADCROP:5). Only the read pairs with both members passing the quality filtering test were considered further (reads passed: 94% in islet, 97% in embryo). The raw reads were mapped on the zebrafish genome (vZv9) using the Tophat2 software (v2.0.8b) (Kim et al., 2013a). A reference gene model file in the Gene Transfer Format (GTF) was used while mapping the reads. The reference GTF file comprised of pooled genomic features from Ensembl genes, mRNA and refgene tracks of the UCSC genome browser for zebrafish (Meyer et al., 2012). In order to define an optimal set of parameters to build the gene models I tested four different mapping strategies (common parameters for all strategies: --mate-inner-dist 223 --mate-std-dev 63 --library-type fr-unstranded --segment-length 21 segment-mismatches 1 –raw-juncs).

- **Stringent**: --no-discordant --no-mixed –prefilter-multihits
- **Relaxed**: --no-mixed –prefilter-multihits
- **NoFilter**: –prefilter-multihits
- **NoFilterNoMulti**: Including all reads mapped more than 20 times on the genome.

The Cufflinks program (v2.1.1) (Trapnell et al., 2010) was used to assemble the reads mapped by using the choosen mapping strategy (parameters: --frag-bias-correct --library-type fr-unstranded --upper-quartile-norm --no-effective-length-

correction). The transcript models generated by Cufflinks for the embryo and islet mappings were merged together by the Cuffcompare utility from the Cufflinks software package (-V -R -r -s -C).

5.2.3 Annotation and differential expression analysis of assembled transcripts

The Annocript pipeline was employed to predict lncRNAs from the assembled transcripts. The differential expression analysis of all the transcripts was performed separately for coding and lncRNA transcripts by a Perl script (run_DE_analysis.pl) from the Trinity software suite (Grabherr et al., 2011) which uses the edgeR (-dispersion 0.1) package (Robinson et al., 2010). Different filtering criteria were used to identify coding and lncRNA transcripts overexpressed in the islet in comparison to whole embryo (Coding: fold change > 2, FDR <= 0.01; lncRNA: fold change > 2, FDR <= 0.1).

5.2.4 Mapping of assembled transcripts with zebrafish Refseq genes and comparison with type 2 diabetes associated genes

The coordinates of the zebrafish Refseq genes were obtained in GTF format from the UCSC database (http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/database/refGene.txt.gz). The assembled transcripts were compared with the Refseq genes as reference, using the Cuffcompare utility from the Cufflinks software package (parameters: -r -R -V -C). The zebrafish transcripts which were successfully mapped to Refseq genes, were assigned the corresponding Refseq gene symbol. List of ~500 coding genes implicated in T2DM were downloaded from the Type 2 Diabetes Genetic

Association Database (T2DGADB) (Lim et al., 2010). The T2DGADB gene symbols were compared with the Refseq gene symbols assigned to zebrafish transcripts by a custom Perl script. Further the list of T2DGADB genes mapped to a Refseq gene by the Perl script, were manually curated to prepare the final list of zebrafish transcripts associated with T2DM.

5.2.5 Detection of sequence conservation and visualisation in the genome browser

The Genome wide phastCons sequence conservation scores for zebrafish were downloaded in BigWig format from the UCSC genome browser database (http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/phastCons8way/vertebrate.phastCons8way.bw). The mean phastCons conservation scores for coding, long non-coding and random intergenic regions were calculated using the bigWigSummary utility from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/). The output files from Tophat2 in BAM format were converted to BigWig format using the genomeCoverageBed binary from the BEDTools package (v2.17) (Quinlan and Hall, 2010) and the bedGraphToBigWig utility from the UCSC database (http://hgdownload.cse.ucsc.edu/admin/exe/). The visualisation of the RNAseq peaks and transcript models was carried out in the Integrative Genomics Viewer (v2.2.7) (Thorvaldsdóttir et al., 2012).

5.2.6 Identification of microsynteny, prediction of alternative polyadenylated transcripts and gene ontology enrichment

The SynLinc pipeline was used to predict putative islet specific microsyntenic lincRNAs between human and zebrafish. The distance of lincRNAs from their closest coding gene was calculated by the closestBed utility from the BEDTools software suite (v2.17) (Quinlan and Hall, 2010). All multiexonic lincRNAs lying within 10 KB of the 3' end of their closest coding gene and transcribed in the same strand were classified as putative alternative polyadenylated transcripts. The classification was repeated for all uniexonic lincRNAs lying within 10 KB of the 3' end of a coding gene without any preference for strand orientation. The gene ontology (The Gene Ontology Consortium, 2012) enrichment analysis was performed on the GO mapping done by the Annocript pipeline using a custom R script exploiting the Fisher exact text and P value FDR correction to select significantly enriched GO classes (minimum representative for a GO class: 5; FDR <= 0.05).

5.3 Results and Discussion

5.3.1 Standardisation of short read mapping for downstream assembly of lincRNAs

5.3.1.1 Issues in mapping of short sequencing reads on the genome

Approximately 42 million short reads from whole embryo and 37 million from islets were generated by sequencing of the RNA samples. More than 90% of reads

from both whole embryo and islet cells passed the quality filtering tests (Embryo: 39116372; Islet: 35718979). Two factors govern the identification of lincRNAs from RNAseq data, the mapping of reads on the genome and the assembly of reads into transcript models. The mapping of raw RNAseq reads on the genome is an arduous task because of the short length of the reads and their extensive multi-mapping. Since many reads fall inside repetitive regions of the genome or map to exons of genes with multiple paralogs it is not surprising that a short read alignment program may place them on multiple locations within the same genome. Further non-coding sequences are known to originate from repeat elements (Smit, 1999) and a recent published report associated transposable elements with the evolution of mammalian lncRNAs (Kelley and Rinn, 2012). It is important to note that amongst all sequenced vertebrate genomes, the zebrafish contains the largest percentage of repeat elements (52.2%, Howe et al., 2013). It is thus imperative to develop a method to map and assemble short sequencing reads which accounts for the repetitive sequences without compromising on lncRNA identification. A previous report mentions that multi mapping of short reads at a maximum given threshold (10 per read) helps in estimating the transcript abundance by allowing the inclusion of homologous sites (Odawara et al., 2011). Another study describes a method to assign multi-mapping reads by distributing them according to the coverage ratios of uniquely mapped reads in each loci (Mortazavi et al., 2008). Popular mapping algorithms like the Tophat2 (Kim et al., 2013a), Bowtie (Langmead and Salzberg, 2012) and Star Aligner (Dobin et al., 2013) support the filtering of multi mapping-reads through various parameters. The

question arises about an ideal set of parameters which may be used to assign multi mapping reads for a lincRNA discovery pipeline. Another issue which plagues the reference based computational transcript assembly is to define the splice sites of unknown or unannotated lowly expressed transcripts (potential lncRNAs) which often do not have a reference gene model and are assembled with the support of few aligned reads. It is important to consider the class of reads used to assemble such transcripts especially if we are interested in the prediction of lncRNAs. Short reads mapped on the genome can fall under three classes

- **Concordant**: both reads of a pair are mapped in correct orientation and separated by expected distance on the same genomic locus.
- **Discordant**: both read pairs are mapped on the same genomic locus but not in correct orientation or not within expected distance threshold.
- **Mixed**: each read of a pair is mapped to a different genomic locus.

The discordant and mixed class of reads can potentially identify structural variants in a genome (Medvedev et al., 2009) but cannot define the splicing pattern of a gene with confidence. This problem is magnified in case of lowly expressed transcript loci without a reference gene model if the majority of mapped reads belong to the discordant and mixed category.

5.3.1.2 Different strategies to map sequencing reads from the islet and embryo samples on the zebrafish genome

The Tophat2 software is one of the most widely used programs for mapping raw sequencing reads on a genome (Kim et al., 2013a) and along with its previous

version (Trapnell et al., 2009) has been cited in over 700 publications. It uses the bowtie program (Langmead and Salzberg, 2012; Langmead et al., 2009) to align the short reads and creates a splice junction database. The reads are then realigned to the splice junction database to verify the first mapping and generate an alignment files which can be used by downstream transcript assembly programs. I performed several mappings of the raw reads from islet cells and whole embryo RNAseq data using Tophat2. Each run was based on specific mapping parameters of the Tophat2 program. The purpose of these mappings was to define an ideal set of Tophat2 parameters which can help in the downstream assembly of putative lincRNA transcripts with high sensitivity. Before explaining the mapping approaches it is important to elaborate on the parameters and terms used in Tophat2 mapping. Tophat2 has two parameters to filter multi-mapped reads (-g and -prefilter-multihits) which work on different principles. The -g parameter allows the end-user to provide a threshold n, based upon which it allows a maximum number of n mappings for all reads. The prefilter-multihits parameter percolates only those reads which have less than n number of mappings. Hence while -g and prefilter-multihits use n as their threshold their approaches differ as -g allows all mapped reads considering the best alignments for each read while -prefilter-multihits removes reads beyond a threshold number of multi-mappings. I have used only the -prefilter-multihits parameter during my analysis. Further the program assigns a score to each genomic alignment of a multi-mapped read in a hierarchical order from the best to the worst alignments. If all alignments are of equal score then the hierarchy is assigned randomly. The best alignment of a short read on the genome

is considered to be the primary alignment while the rest are labeled as secondary alignments. Keeping into account the manner in which Tophat2 treats multi-mapped reads, four mapping approaches were defined:

- **Stringent**: Only concordant reads are mapped to the genome with removal of reads mapped more than 20 times on the genome (--prefilter-multihits 20).
- **Relaxed**: Both concordant and discordant reads are mapped to the genome with removal of reads mapped more than 20 times on the genome (--prefilter-multihits 20).
- **NoFilter**: Concordant, discordant and mixed reads are mapped on the genome with removal of reads mapped more than 20 times on the genome (--prefilter-multihits 20).
- **NoFilterNoMulti**: All reads are mapped to the genome, the default Tophat2 parameters.

The mapped reads from whole embryo and islet cells were used separately to build transcript models. The transcript models generated for islet and embryo were merged together to generate a final transcript dataset. The Cufflinks pipeline encourages the use of Cuffmerge utility to merge the transcripts from different assemblies. Cuffmerge performs its own hard coded assembly on transcripts from each sample. Which basically means that Cuffmerge breaks the assembled transcript models into short fragments, pools them together and re-aligns them on the genome followed by another Cufflinks assembly to generate a merged

transcript dataset. I got an inflated number of transcripts on using Cuffmerge with almost 25% of reported loci not having a 100% read coverage. This happened because many transcripts were being reported just because they are present in the reference GTF file provided even though they do not have mapped read support. The inflated numbers by Cuffmerge led me to test transcript merging with Cuffcompare (another utility in the Cufflinks software suite). The Cuffcompare utility is more stringent in its approach to merge transcripts since it compares the splice junctions of the assembled transcripts to calculate their probability of being isoforms of a same gene/loci. Also while comparing against a reference transcript file, Cuffcompare only reports those transcript models which have support from mapped reads. The following example shows the difference between the Cuffmerge and Cuffcompare approach. There are two transcripts, A and B, each with a couple of exons. If A and B overlap, and they don't disagree on splicing structure, they could possibly belong to the same gene. Cuffcompare will only merge them if A is "contained" in B, or vice versa. That is, only if one of the transfrags is essentially redundant. Otherwise, they both get included. Cuffmerge on the other hand, will merge them if they overlap, and agree on splicing, and are in the same orientiation. This behavior by Cuffmerge may also lead to spurious transcript models which may mislead downstream experimental validations. Hence I decided to consider the merged transcript models generated by Cuffcompare. Maximum number of transcripts were assembled by the NoFilterNoMulti approach while interestingly the Relaxed approach generated the least number of transcripts **(Table 5.1)**.

Type of mapping	Number of transcripts	Number of loci
Stringent	36921	20924
Relaxed	36502	20684
NoFilter	42855	24878
NoFilterNoMulti	59747	35555

Table 5.1 Count of transcripts generated by the different mapping approaches.

Based on the conservativeness of mapping I expected the *Stringent* mapping to generate the minimum number of transcripts. On close inspection of a few transcript loci present in *Stringent* mapping and absent in *Relaxed* I found the presence of additional reads with secondary alignments in the *Relaxed* category. The presence of reads with secondary alignments above a threshold percentage (unknown hard coded parameter of Cufflinks) leads to no transcript generation in a genomic locus by Cufflinks. Further I compared the transcript models generated by each approach with the reference Ensembl coding and long non-coding gene models for zebrafish (**Figure 5.2**). There is an increase in number of reference lncRNAs overlapping predicted transcript models in the NoFilter and NoFilterNoMulti assemblies in comparison with *Stringent* and *Relaxed* (**Figure 5.2 A**). The number of overlapping coding genes remains almost constant for all categories. Yet, in terms of accuracy, the *Stringent* and *Relaxed* mapping show the highest sensitivity in identifying known exon models of reference coding and lncRNA genes (**Figure 5.2 B**). Therefore, although the less stringent approaches are able to predict a higher number of transcripts they result less accurate. The difference in accuracy of the predicted exon models is more prominent in case of lncRNAs for the *Stringent* and *Relaxed* approaches. Since lncRNAs are reported to

be expressed at a lower level in comparison to coding genes (Pauli et al., 2011a; Ulitsky et al., 2011) this observation suggests that the change in mapping parameters may affect the assembly of lowly expressed non-coding transcripts. Often such transcripts fall in intergenic regions and overlap repetitive regions. The lack of proper paired reads mapping on such region produces transcript models based solely on unpaired or multi-mapped reads. Hence it is difficult to demonstrate the verity of the transcript structure and its expression. I selected two specific examples of a coding and a non-coding region to highlight the differences in transcript assembly from the different mapping outputs.

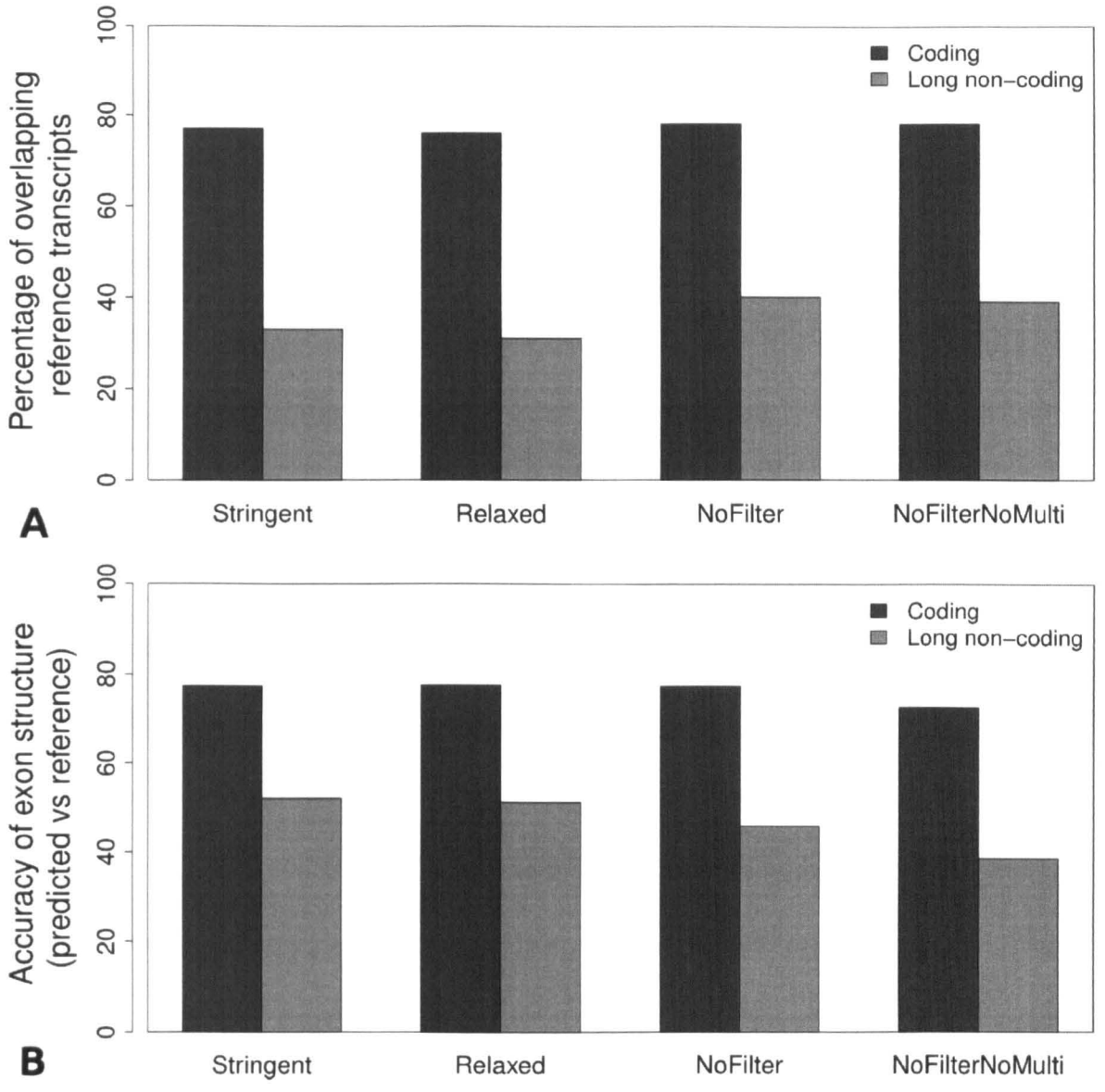


Figure 5.2 Accuracy of Tophat2 mapping **A)** Number of reference transcripts from Ensembl (v73) overlapping predicted gene models **B)** The sensitivity of exon structure in the predicted transcript models which overlap with the reference transcript models, defined as number of correctly predicted exons/total number of reference exons.

5.3.1.3 Example of assembled transcripts demonstrating the differences in different short sequencing read mapping strategies

The first example is of an intergenic region where islet specific transcription occurs (**Figure 5.3**). This region is devoid of an annotated transcript feature (coding/non-coding) but shows the presence of transcription. The *NoFilterNoMulti* approach gives a long stretch of mapped reads which result in a long 3' exon of the transcript model (*NFNM_00054304*). This exon overlaps repetitive regions on the genome and is comprised of multi-mapped reads. The *Relaxed* and the *NoFilter* approaches assemble shorter 3' exons (*RELX_00033050, NOFIL_00039045*) but they differ between the position of the 3' end splice junction due to presence of mixed and discordant mapping. Additional single exonic transcript models are also reported in *NoFilter* (*NOFIL_00039247, NOFIL_00039248*). Even though reads from the islet transcriptome are mapped, Cufflinks does not generate a transcript model for the *Stringent* approach. It must be noted that while the *Stringent* mapping considers only concordant read pairs aligned to the genome it also allows multi mapping up to 20 times genome-wide (default by Tophat2). Since only reads with secondary alignments of concordant reads define the putative splice junctions of the transcript model in this example, it is not considered by the Cufflinks program in case of *Stringent* mapping.

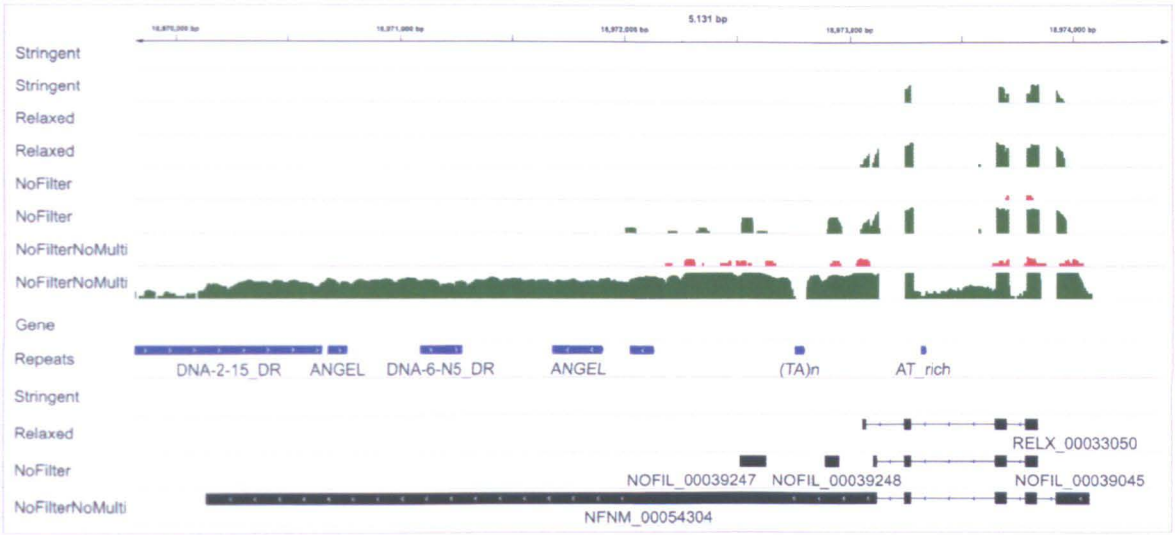


Figure 5.3 Genome browser screenshot of an intergenic region (chr4:18,972,103-18,974,896) on the zebrafish genome. The Gene track represents Refseq gene models and has no genes overlapping the given region. The tracks above Gene are the coverage of reads mapped on the genome from islet (green) and 72 hpf whole embryo (red) using various Tophat2 mapping approaches. The tracks below Gene represent the transcripts assembled by pooling the reads from islet and embryo for each mapping approach.

The second example is the coding gene Slit homolog 1 (*Slit1B*). The *Slit* family genes (*Slit1, Slit2, Slit3*) are implicated in protection of islet cells from apoptosis and regulation of insulin secretion (Yang et al., 2013c). They also play a role in the axon guidance during development of dopaminergic neurons (Cornide-Petronio and Barreiro-Iglesias, 2013). It is interesting to note that the *SLIT1B* gene model as generated from the *Stringent* mapping differs from the other transcript models (*TCONS_00024365*) (**Figure 5.4**). The other mapping approaches show the presence of an additional 5' exon overlapping a repeat region (*RELX_00024237, NOFIL_0029604, NFNM_00039364*) generated exclusively due to multi-mapped

reads. These observations indicate the importance of parameter choices which can inordinately alter the structure and expression abundance of predicted transcripts. I chose the results from the *Stringent* mapping as my final transcript dataset because of the following reasons:

- The assembled transcripts are built from properly paired reads.
- The assembly has better accuracy in defining intron/exon boundaries.
- The estimation of expression from mutli-mapped reads is minimal.

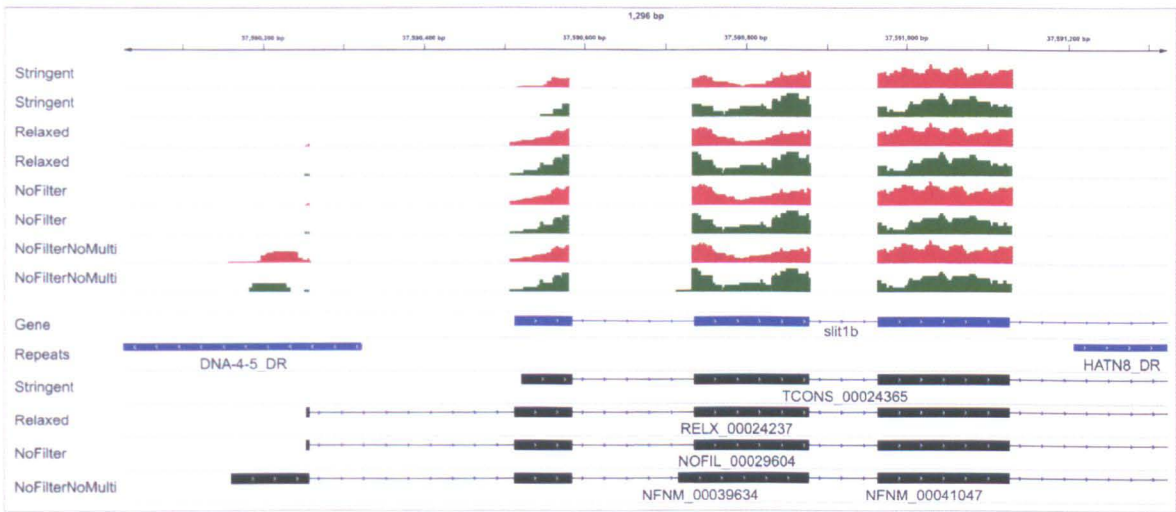


Figure 5.4 Genome browser screenshot of the first three exons of the *Slit1b* gene (chr22:37,589,891-37,591,373) on the zebrafish genome. The Gene track represents Refseq gene models. The tracks above Gene are the coverage of reads mapped on the genome from islet (green) and 72 hpf whole embryo (red) using various Tophat2 mapping approaches. The tracks below Gene represent the transcripts assembled by pooling the reads from islet and embryo for each mapping approach.

5.3.2 Annotation of the assembled transcripts and prediction of long non-coding RNAs

In total ~37,000 transcripts were assembled using the stringent approach which

were systematically annotated, categorized and manually curated to identify candidate lincRNAs differentially expressed in islet cells (**Figure 5.5**). The Annocript pipeline was employed to annotate and predict 35,110 coding transcripts and 227 lncRNAs using default parameters of lncRNA prediction. Around ~1500 transcripts were classified as unknown since they were without annotation but could not be classified as an lncRNA due to constraint of ORF size (<= 100 AAs) and non-coding potential score (>= 0.95). I used the previously developed Annocript pipeline to predict the putative lncRNA sequences. As per a prior observation the pipeline relies on the Portrait Non-Coding Potential (NCP) score for all Potential Long Non-Coding sequences (PLoNCs: sequences with no annotation and ORF < 100 AAs) to predict the final set of lncRNAs (Arrial et al., 2009).

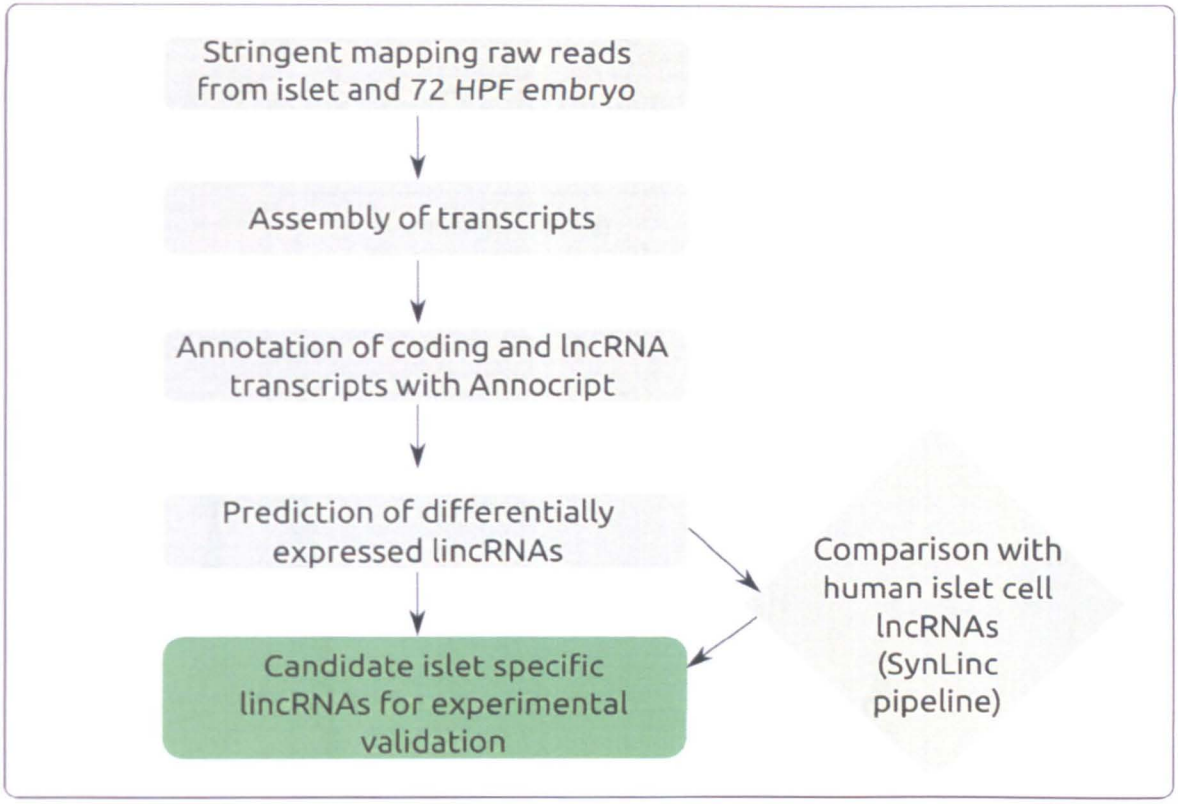


Figure 5.5 Pipeline for identification of differentially expressed lincRNAs in zebrafish islet cells.

A conservative NCP threshold may prove to be a good strategy in case of *de novo* transcriptomes with no reference genome or to compare against coding genes in order to limit the number of false positives, but have the drawback of producing a high number of false negatives. As already mentioned, the Portrait authors suggest that a cutoff of 0.5 can be accepted to distinguish between conding and non-coding transcripts with acceptable confidence. In my specific case I also have to take into account that the gene models were already built with high stringency, therefore I decided to plot the distribution of NCP scores segregating the coding sequences and PLoNCs based on the Annocript results (**Figure 5.6**). A NCP score of 0.5 or above is observed to be suitable for separating the coding sequences from the

PLoNCs. However I chose to be more stringent and considered the mean NCP of the PLoNCs (0.72) as my cut-off for the prediction of putative lncRNAs.

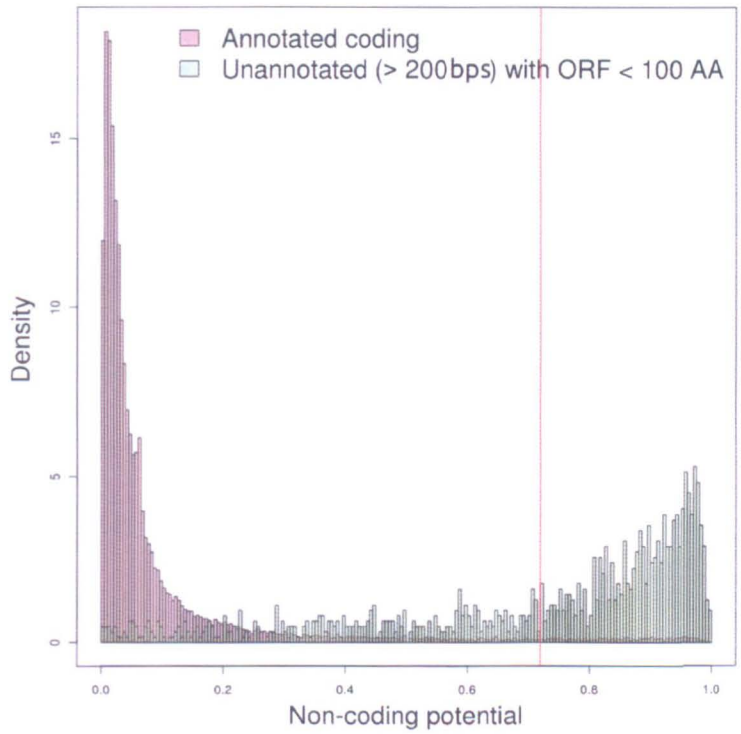


Figure 5.6 The non-coding potential score distribution for coding and potential lncRNA sequences annotated by the Annocript pipeline. The vertical red line marks the mean score of non-coding potential (0.72) for the potential lncRNA sequences (PLoNCs). The x-axis represents the non-coding potential (NCP) score assigned to the lincRNAs. The y-axis represents the frequency of the lincRNAs at a given NCP score. The bars in pink represent those lincRNAs which are predicted to be coding by Annocript, while the green bars represent the lincRNAs predicted to be Potential Long Non-Coding by Annocript.

Interestingly, this cutoff represent the point in which the slope of the distribution of the NCP scores start to increase suggesting a better classification capability of Portrait and a lower potential number of false positives and false negatives. The new cut-off score led to the prediction of 805 lncRNAs of which 178 are lincRNAs

(long intergenic noncoding RNAs).

5.3.3 Identification of assembled transcripts differentially up-regulated in the islet cells

The purpose of the RNAseq experiment is to find the putative lincRNA candidates in zebrafish which may play a role in the pancreatic metabolism, specially in the development and differentiation of islet cells. Such an implication may probably associate the transcription of a lincRNA with the prognosis of a pancreatic disorder like the T2DM. Hence I decided to find the assembled transcripts which are expressed at a higher level in the islet cells as compared to the whole embryo. I used the Bioconductor edgeR package (Robinson et al., 2010) to find all differentially expressed transcripts in the assembled transcriptome. The edgeR package generates an over-dispersed Poisson model with the raw count of reads representing each transcript to estimate the expression variability (Robinson and Smyth, 2007). The edgeR program is intended to be used for data with biological replicates. It calculates a dispersion value for each sample from the biological replicates, which is a measure of the variability of expression measures within samples. Since the samples I worked with do not have biological replicates I used a dispersion value of 0.1 as suggested by the edgeR software manual (http://www.bioconductor.org/packages/2.13/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf) (Robinson et al., 2010). The RNAseq experiment were carried out with samples from whole embryo and islet cells at the same stage of development, hence the transcripts expressed in the islets may be also detected in the whole embryo. This situation may result in a smaller

variation between the expression of transcripts in both the samples. I expected the variation to be far less pronounced in lncRNAs as compared to coding genes they are lowly expressed. In fact the mean expression of coding transcripts in both whole embryo and islet cells is 3X the expression of lncRNA transcripts. The larger number and higher expression level of coding genes may influence the detection of small change of expression in the lncRNAs. Hence I performed the differential expression analysis on the coding and the lncRNAs transcripts separately using different cutoffs (coding: fold change > 2, FDR <= 0.01; lncRNA: fold change > 2, FDR <= 0.1). I identified 939 coding transcripts and 94 lncRNAs to be significantly overexpressed in the islet cells.

5.3.4 Gene ontology enrichment analysis of coding transcripts predicted to be differentially up-regulated in islet cells

Further I wanted to check if the coding transcripts with elevated expression levels in islet cells could be associated with a potential function in pancreatic development and metabolism. Hence I performed a gene ontology enrichment analysis for the differentially expressed coding genes in the islet cells (Minimum number of genes = 5; P value corrected < 0.05) (**Figure 5.7**). Amongst the GO terms which have been significantly enriched are *"cellular glucose homeostatis"* and *"type B pancreatic cell differentiation"* which indicate an enrichment of genes pertaining to pancreatic development and metabolism. The term *"proteolysis"* is also amongst the enriched terms. Proteolysis is defined as *"the hydrolysis of proteins into smaller polypeptides and/or amino acids by cleavage of their peptide bonds"* (Ashburner et al.,

2000). The islet cells of the pancreas are involved in the secretion of hormones like insulin, somatostatin and glucagon which involve several signaling and degradation pathways. A few recent reports have emphasised on the selective degradation and post-translational modification of specific proteins to aid in the functioning of pancreatic cells (Aston-Mourney et al., 2013; Chen et al., 2013b; Tiwari et al., 2013). Hence the enrichment of the term *"proteolysis"* is in agreement with the potential up-regulation of various coding genes involved in proteolytic activities within the islet cells.

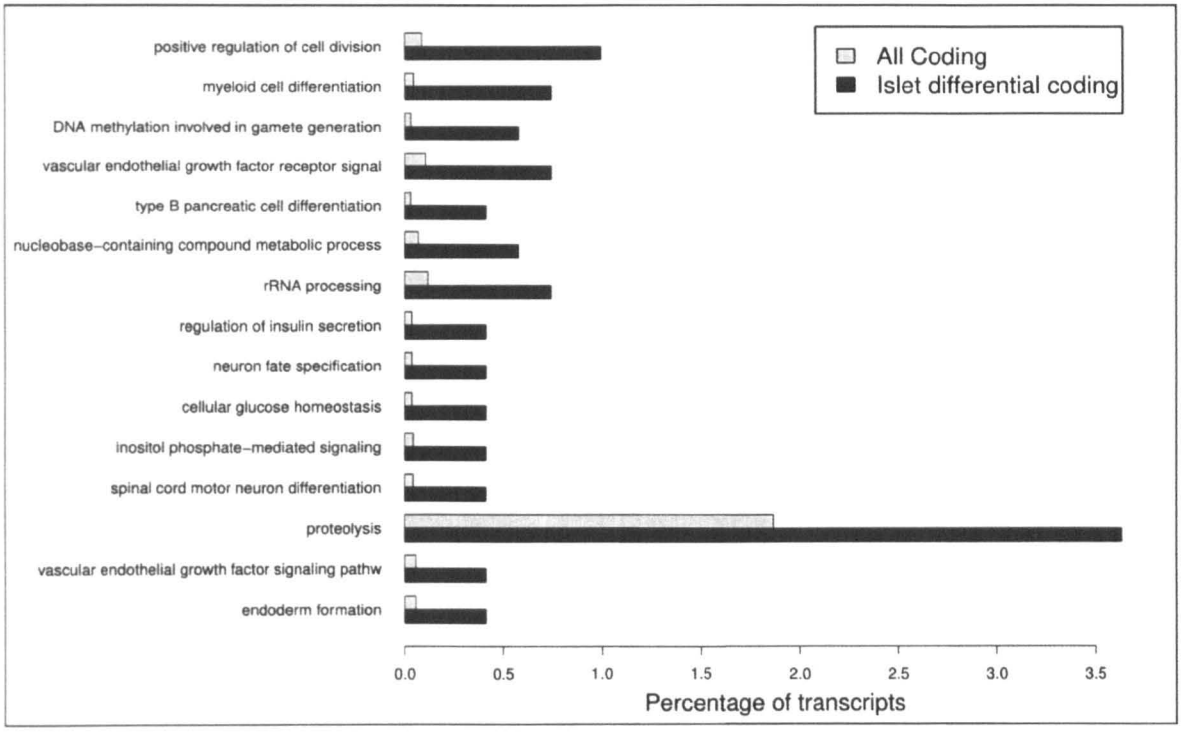


Figure 5.7 Gene ontology enrichment analysis for differentially overexpressed coding genes in the zebrafish pancreatic islet cells. The x-axis represents the percentage of transcripts which are defined by a particular GO biological processs. The y-axis represents the gene ontology biological process terms.

It is also interesting to find the GO terms like *"neuron fate specification"*, *"spinal cord and motor neuron differentiation"* and *"neuropeptide signaling"* enriched in the

differentially expressed coding genes. In fact pancreatic and neuronal cells share a common evolutionary origin and neuronal cells of coelentrates are reported to evolve into neuroendocrinal cells of invertebrates which further diversify into neuronal and endocrine cells in higher vertebrates (Falkmer, 1993). A classic example of such diversification is the secretion of insulin, in the neurons of invertebrates, in visceral endocrine cells in chordates and a very low but detectable insulin secretion in mammalian neural cells (Devaskar et al., 1994). Further the β-cells in vertebrates are reported to communicate with the hyptohalamus through a signaling pathway to regulate the secretion of insulin (Gelling et al., 2006). Also a recent report suggests a high level of similarity in epigenetic modifications marking an active chromatin state, between pancreatic *beta* cells and neuronal tissues in mouse (van Arensbergen et al., 2010). These findings suggest the presence of shared or common signaling pathways between neuronal and pancreatic differentiating cells which are probably reflected by the enrichment of specific GO terms related to nervous system development.

5.3.5 Association of differentially up-regulated coding genes with type 2 diabetes

In order to associate the differentially expressed coding genes with a potential role in diabetes I mapped the predicted coding transcripts with the reference set of zebrafish Refseq coding genes (with known gene function) to obtain a putative Refseq gene IDs and symbols for the assembled transcripts. I could assign a Refseq gene ID to 50.6% (10,945 genes for 17,772 transcripts) of all coding transcripts and

to 60% of all differentially expressed transcripts (372 genes for 504 transcripts). Further I compared gene symbols associated with type 2 diabetes from the previously published type 2 diabetes genetic association database (T2DGADB; 530 genes) (Lim et al., 2010) to the mapped Refseq genes symbols using a custom Perl script. The script predicted 301 T2DGADB gene symbols to match with 489 Refseq gene symbols. I further curated the results manually to select 289 T2GADB genes mapped to 412 Refseq genes in zebrafish. I found a significantly higher number of Refseq genes to be diabetes associated amongst the differentially expressed subset (6.9%, 26 out of 372) in comparison to the whole transcriptome (3.7%, 412 out of 10,945; two sample proportion test, P value: 0.002). This indicates that the transcripts expressed in zebrafish islet cells are enriched for homologs of human genes associated with type 2 diabetes which play a crucial role in pancreatic cell development and differentiation (**Table 5.2**) including the well known *Isl1* and *Ins* genes.

Gene symbol	Name	Function	Reference
Ins	Insulin gene	Regulation of glucose metabolism	(Bell et al., 1980)
Isl1	Islet 1	Formation of exocrine and endocrine pancreatic tissues	(Wilfinger et al., 2013)
PDX1	Pancreatic and duodenal homeobox 1	Pancreatic lineage development	(Liang et al., 2013)
CDK5	Cyclin-dependent kinase 5	Regulation of epidermal growth factor dependent insulin secretion	(Lee et al., 2008a)
GCK	Glucokinase	Pancreatic glucose sensor	(Matschinsky, 2002)
FoxA2	Forkhead box protein A2	Maintenence of beta cell metabolic pathways	(Gao et al., 2010)
ABCC8	ATP-binding cassette, sub-family C (CFTR/MRP), member 8	Regulation of pancreatic cell excitability and insulin secretion	(Karaskov et al., 2006)
HNF1a	Hepatic Nuclear Factor 1	Early development of pancreas	(Haumaitre et al., 2005)
SLC2A2	Solute carrier family 2	Regulation of metabolic pathways guiding insulin secretion	(Matschinsky, 2002)
PAX4	Paired box gene 4	Differentiation of progentior cells into alpha and beta cells	(Collombat et al., 2009)

Table 5.2 A list of coding genes functionally important in pancreatic disease and development which were predicted to be differentially expressed in zebrafish islet cells.

Hence the differential expression of such genes in the zebrafish islet transcriptome suggests potential common pathways and regulatory mechanisms, which can be studied in the fish model system in an early developmental stage.

5.3.6 Structural features of the predicted coding and long non-coding transcripts

To understand a possible role of lncRNAs pertaining to islet cells development and differentiation I decided to focus on the lncRNAs predicted by the transcript annotation. In order to understand the structural features of the predicted

lncRNAs, I compared them with a random sample of coding transcripts (**Figure 5.8**). The lncRNAs are observed to be smaller in length with fewer but longer exons in comparison to the coding transcripts. The fewer and longer exons are due to majority of the predicted lncRNAs being monoexonic (75%) which is similar to previously reported human islet cell lncRNAs (74%) (Morán et al., 2012). The fewer number of exons also explains the smaller length in comparison to coding genes. In fact long non-coding RNAs are known to be spliced inefficiently often leading to mono or bi exonic transcripts (Tilgner et al., 2012). This fact is corroborated by the GENCODE catalog of lncRNAs which are reported to be biased towards having a single canonical splice site within the transcript (Derrien et al., 2012). The lack of an efficient splicing mechanism may stem from two different reasons. Firstly, a lncRNA can influence the expression of its proximal coding gene by simply the act of its transcription rather than by an activity exploited by a specific transcriptional product as exemplified by the action of Antisense of IGF2R non-protein coding RNA (*AIRN*) (Latos et al., 2012). Secondly, a splicing independent, locus specific mechanism of the lncRNA may result in the interaction of the lncRNA molecule with a protein to influence the expression or chromatin state of its genomic neighborhood. A good example is the Hoxa distal transcript antisense RNA (*HOTTIP*) (Burgess, 2011) lncRNA which lies at the tip of the Homeobox A (*HoxA*) cluster genes and by chromosomal looping influences the expression of the *HoxA* genes (Wang et al., 2011).

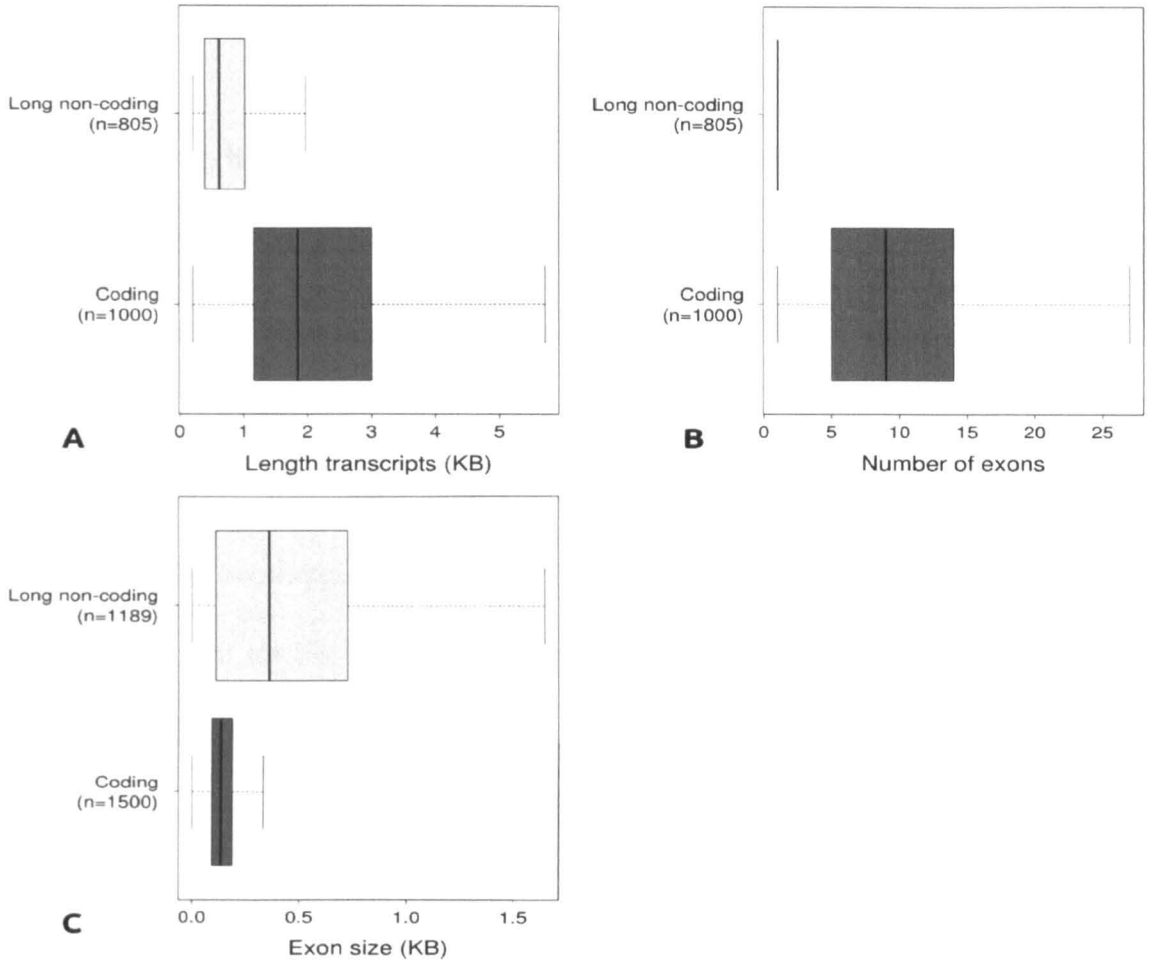


Figure 5.8 Structural features of coding and long non-coding transcripts **A)** Length of transcripts **B)** Number of exons **C)** Size of exons.

5.3.7 Conservation of sequence in the predicted coding and long non-coding transcripts

Between the coding and the long non-coding transcripts the lincRNAs have the least sequence conservation, even lower than random intergenic regions (**Figure 5.9**). Thus the low level of lincRNA sequence conservation reflects the high rate of evolutionary turnover of these sequences. A large proportion of vertebrate lncRNAs co-occur with transposable elements (TEs) with significant variances in the class of TEs overlapping lncRNAs in different species (Kapusta et al., 2013). The TEs may account for the high rate of sequence diversification of lncRNAs. While a subset of lncRNAs overlap conserved genomic regions of coding gene exons thus partially retain their sequence, the lincRNAs are completely intergenic and hence might be under a lower selective pressure for sequence conservation.

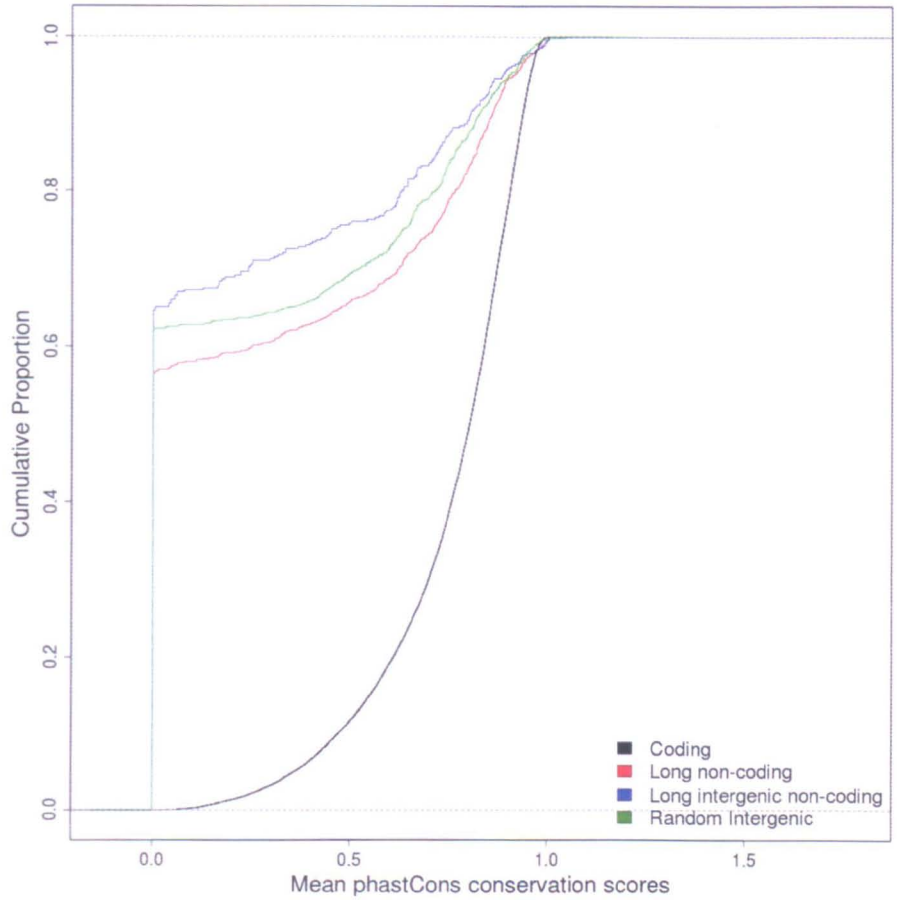


Figure 5.9 Mean phastCons 8 way conservation scores of zebrafish for coding, long non-coding and long intergenic non-coding transcripts. The conservation scores are calculated by aligning the zebrafish genome with the human, mouse, *X. tropicalis*, *Tetraodon*, medaka, fugu and stickleback. The x-axis represents mean phastCons conservation scores and the y-axis represents the cumulative proportion of transcripts.

5.3.8 Expression abundance of the coding and the long non-coding transcripts in whole embryo and islet cells

I compared the expression pattern of the predicted coding and lncRNA transcripts. As expected, I observed that the coding transcripts are expressed at a higher level as compared to lncRNAs (**Figure 5.10**). Interestingly, the average expression of

lincRNAs is higher in comparison to lncRNAs (**Figure 5.10 C, D**) and the difference is quite prominent for the islet cells. This observation points towards an enrichment of predicted lincRNAs expression in the islet cells. Hence, I compared the percentage of coding, lncRNA and lincRNA transcripts differentially expressed in the islet cells with total number of transcripts in each class originating from the islet cells (supported by the assembly of transcripts from islet cells by Cufflinks). It is interesting to note that a significant percentage of differentially expressed lncRNAs (9%; p.val: 0.006) tend to lie near a differentially expressed coding gene (10 kb distance threshold) as compared to all predicted lncRNAs (3%).

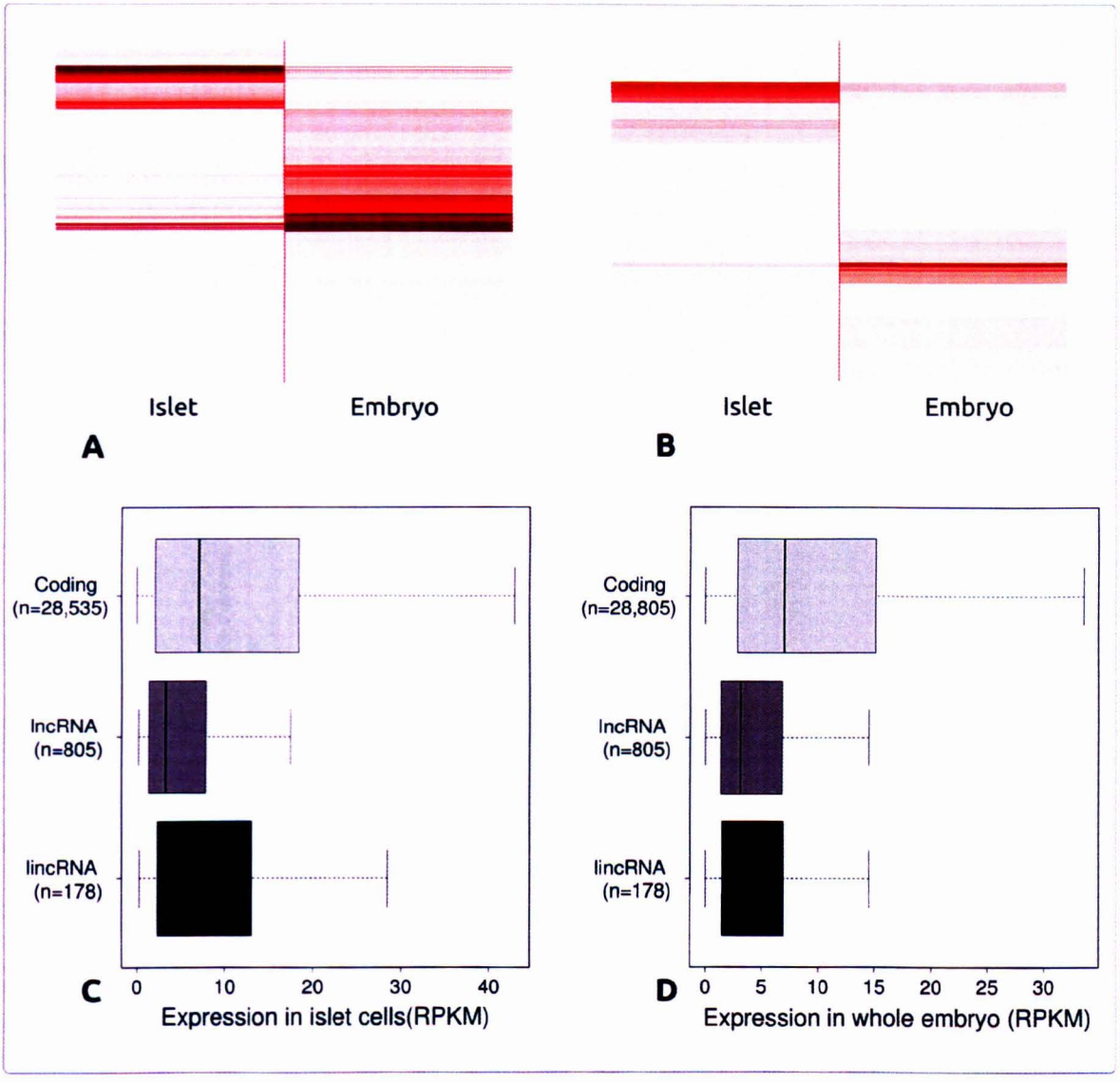


Figure 5.10 Heatmap of expression level for differentially expressed transcripts in both islet cells and whole embryo **A)** Coding **B)** Long non-coding. Distribution of transcriptional abundance of all predicted transcripts (coding, long non-coding and long intergenic non-coding) in **C)** Islet cells **D)** Whole embryo, at 72 hours post-fertilization.

In addition, I found a significantly higher number of lincRNAs are differentially expressed in the islet cells as compared to lncRNAs and coding transcripts (**Figure 5.11**). LincRNAs are known to be lowly expressed and highly tissue and cell type specific (Aprea et al., 2013; Cabili et al., 2011; Mercer et al., 2010; Pang et al., 2009).

Therefore the sequencing of a very specific cell population, the islet, and a more general whole embryo sample, permitted to have enough reads support for islets specific lincRNAs while diluting other embryonic cell specific transcripts.

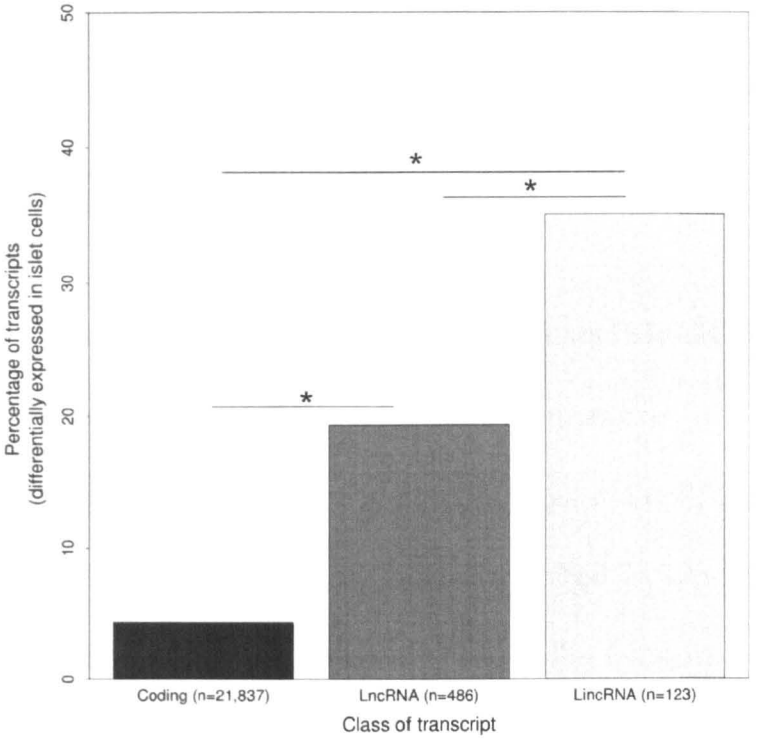


Figure 5.11 Percentage of differentially expressed transcripts in coding and long non-coding categories

5.3.9 Selection of candidate lincRNAs for experimental validation

Hence I performed a manual curation of the predicted lincRNAs to select candidates for experimental validation based on the following criteria

- Differential expression of the lincRNAs in islet cells.
- Presence of splicing. The preference for multi exonic lincRNAs was due to the fact that their splice sites can be easily targeted by morpholino based knock down studies.

- Low probability of the lincRNA being an alternative polyadenylation event of a coding gene . All multiexonic lincRNAs lying within 10 KB of 3' end of a flanking coding gene and transcribed in the same strand were classified as putative alternative polyadenylated transcripts (Miura et al., 2013). The classification was repeated for uniexonic lincRNAs without any preference for strand orientation with the same distance threshold.

Based on the manual inspection I selected 26 multi exonic lincRNA transcripts for possible downstream experimental validation (**Annexure 2**). Within these 26 candidates I searched for those which are the closest lincRNAs (within 1 MB upstream or downstream) to a coding gene implicated in type 2 diabetes (from T2DGADB) and differentially expressed in islet cells. I found a single lincRNA transcript, *linc_dvl3* (TCONS_00019413) lying at a distance of 380 kb upstream from the insulin like 5b (*Insl5B)* gene (**Figure 5.12 A**). This lincRNA lies near the Dishevelled segment polarity protein 3 (*Dvl3*) gene which is an integral component of the Wingless-type MMTV integration site (*Wnt*) signalling pathway (Lee et al., 2008b) but is not differentially expressed in the islet cells. The *Insl5B* gene is reported to be expressed in the hypothalamus and colorectum and is involved in the regulation of insulin secretion and β-cell homeostatis (Burnicka-Turek et al., 2012). The region between the lincRNA and *Insl5B* is interspersed by 10 coding genes none of which are differentially expressed in the islet cells. While the islet specific expression of the coding and long non-coding genes in a genomic loci can be the outcome of shared regulatory mechanisms or common pathways further

experimental evidence is required to postulate a direct influence of the lincRNA on the coding gene. Amongst the other candidates another promising example is the multiexonic lincRNA, *linc_samd11* (TCONS_00026726) (**Figure 5.12 B**). Its closest coding gene is the Sterile Alpha Motif Domain containing 11 (*SAMD11*) (distance 35 KB) which is reported to play a role in the cell proliferation with enriched expression in developing retinal photoreceptors and the adult pineal gland (Inoue et al., 2006). While *SAMD11* is not differentially expressed in the islet cells the *linc_samd11* shows a clear islet specific expression pattern. Even though associating lincRNAs with their closest coding genes has remained a standard practice in recent years (Ulitsky et al., 2011; Volders et al., 2012) it does not necessarily imply a regulatory role of the lincRNA with respect to the coding genes but is a strategy for annotation while the function remains unknown.

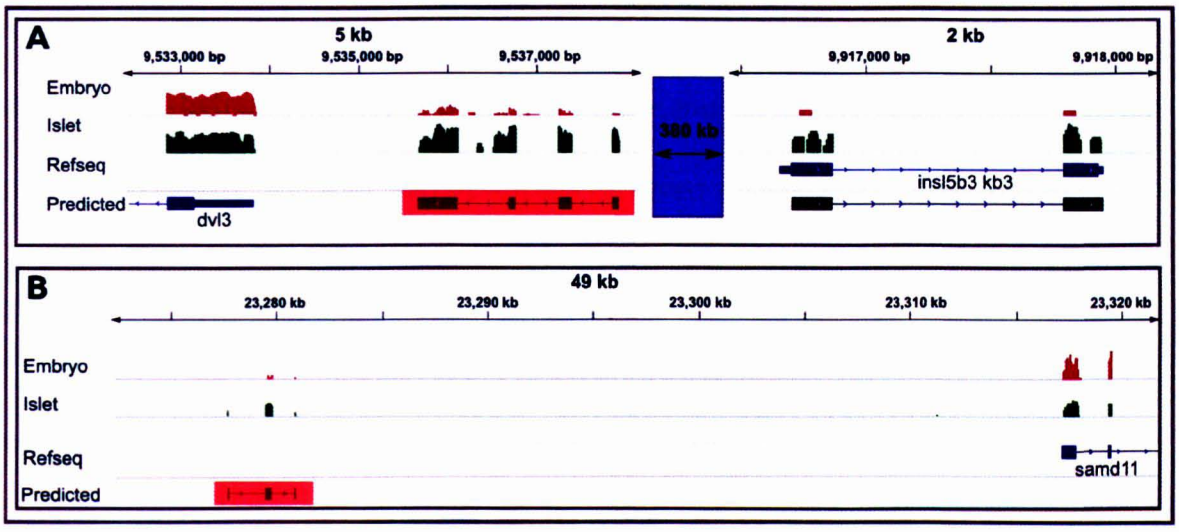


Figure 5.12 Genome browser screen shot of zebrafish lincRNA differentially expressed in islet cells. **A)** *Linc_dvl3* (*TCONS_00019413*) upstream of the *Insl5B* coding gene (chr2:9,502,164-9,971,242) **B)** *Linc_samd11* (*TCONS_00026726*) upstream of the *SAMD11* coding gene (chr23:23,272,151-23,321,834). The Embryo and Islet tracks represent the coverage depth of the mapped RNAseq reads. The Refseq track shows the Refseq genes and the Predicted track represents the gene models generated from the pooled Embryo and Islet transcriptome data. The lincRNA of interest are marked by the red box in the predicted track.

5.3.10 Association of human and zebrafish islet cell lincRNAs by microsynteny analysis

I wanted to check for potential islet specific lincRNAs in zebrafish which may retain their function over long evolutionary distances. Hence I compared previously published human islet cell specific lincRNAs (Morán et al., 2012) with the zebrafish islet cell differential lincRNAs to predict 6 zebrafish lincRNAs paired with 8 human lincRNAs based on the conserved microsynteny of their orthologous flanking coding genes. On manual inspection, for 3 of the 6 fish

lincRNAs, the putative human lincRNA orthologs are associated closely in divergent orientation to the flanking coding gene which was not used to predict the microsynteny. Another two examples involved monoexonic lincRNAs while the last microsyntenic pair had both the human and the zebrafish (TCONS_00021352) orthologs upstream the Ubiquitin Specific Peptidase 46 (*USP46*) gene in divergent orientation (**Figure 5.13**). The *USP46* is a deubiquitinating enzyme which is reported to control the glutamate receptor trafficking in the ventral nerve cord of *C.elegans* (Kowalski et al., 2011). The *USP46* gene is expressed in both islet cells and embryo in the zebrafish and there are no published reports indicating its influence on the pancreatic biology. It is interesting to observe the conserved arrangement of a lincRNA and a coding gene between a large evolutionary distance even though only the lincRNA is enriched to be expressed in islet cells. Yet the few predicted conserved candidates suggest a lack of positional retention of the zebrafish lincRNAs up-regulated in the islet cells in human.

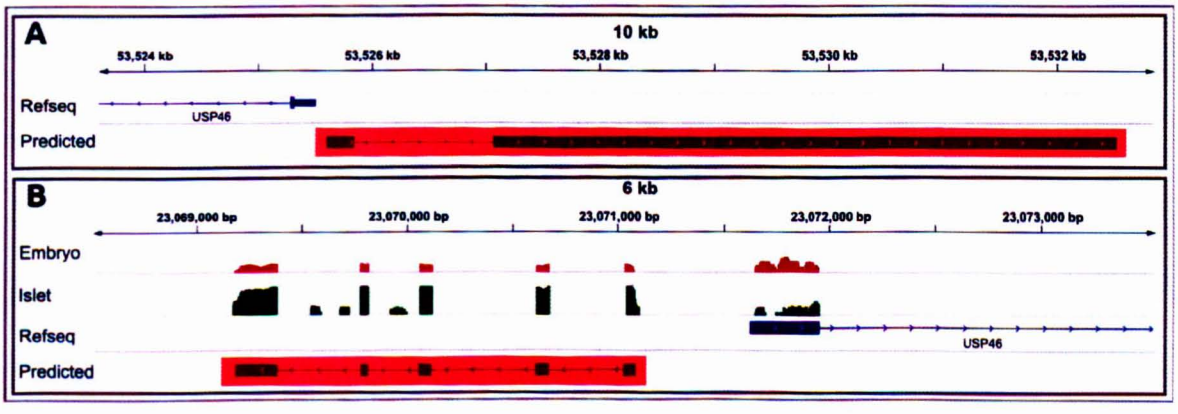


Figure 5.13 Putative conserved lincRNAs with enriched expression in islet cells **A)** Human (*HI-LNC596*:chr5:178,362,381-178,369,686) **B)** Zebrafish (*Linc_usp46* (*TCONS_00021352*): chr20:23,067,628-23,072,955). The lincRNAs are marked by red color boxes.

5.4 Conclusion

The zebrafish islet cell transcriptome was generated with an aim to understand the possible role of lncRNAs in the context of pancreatic biology and their possible role in Type 2 Diabetes Mellitus (T2DM). The choice of mapping strategy for aligning the short reads to the genome has a major influence of the number and structure of lncRNA transcripts assembled downstream. I have defined a specific short read mapping strategy which can assemble lincRNA transcripts with high sensitivity using the current state of the art software programs. Further I used a software pipeline I have developed previously to predict coding and long non-coding transcripts in the islet cell transcriptome. In this study I compared RNAseq data from zebrafish whole embryos and islet cells at 72 hour post fertilization to identify 805 long non-coding RNAs (lncRNAs) expressed in the islet cells of which 94 are predicted to be differentially over expressed in islet cells. The differential lncRNAs tend to lie near coding genes which are themselves differentially expressed in the islet cells. I have developed a novel pipeline for identification of cell type/tissue specific lncRNAs. This pipeline utilizes a specific set of parameters deemed suitable to map and assemble short reads on a genome with minimum ambiguity followed by using a sequence annotation pipeline (Annocript) to predict the coding and long non-coding transcripts. Many of the coding genes predicted to be differentially expressed in the islet cells are implicated in T2DM with well known function in regulation of insulin and differentiation of pancreatic β-cells. The predicted lincRNAs are enriched to be expressed in the islet cells in comparison to all lncRNAs and the coding genes. A few promising candidate

lincRNAs (26 transcripts) were identified based on their differential expression in islet cells, distance from the closest coding gene and splicing pattern. Currently experimental validation of a 15 of these candidate transcripts are being carried out by Irene Miguel-Escalada to verify their expression specificity and observe any phenotypes generated by their knock-down in zebrafish embryos. Thus the current study highlights a strategy to assemble and predict lincRNAs from a transcriptomic dataset and provides the first resource of zebrafish lincRNAs with potential implications in development and differentiation of pancreatic islet cells.

Chapter 6

The early developmental transcriptome of *Tetraodon nigroviridis*

6.1 Introduction

6.1.1 *Tetraodon* as a model to understand the vertebrate embryogenesis

A smaller genome size ($1/8^{th}$ of human genome) and an expected similarity in the gene repertoire were proposed as two major reasons in support of sequencing the fugu genome (Brenner et al., 1993). The sequencing of the fugu (Aparicio et al., 2002) as well as the closely related *Tetraodon* (Jaillon et al., 2004) led to the identification of novel genes in vertebrates as well as further asserted the role of a WGD event in the divergence of the teleost lineage. Compared to the other teleost fishes with sequenced genomes, the *Tetraodon* and fugu have highly compact genomes which proves to be an advantage for comparative genomics studies (Peer, 2004). In fact an estimation of the number of coding genes in human was made based upon homology relationships between human and *Tetraodon* coding genes (Roest Crollius et al., 2000). The *Tetraodon nigroviridis* is a freshwater pufferfish occasionally found in sea water (Jaillon et al., 2004). Amongst all vertebrates the *Tetraodon* has the smallest known genome, characterized by low repeat content, a high gene density and chromosomal stability (Jaillon et al., 2004). The majority of

genes originating from a single chromosome in *Tetraodon* tend to have their paralogs in a single other chromosome and comparison with the human genome showed extensive gene duplication in *Tetraodon* (Jaillon et al., 2004). In the current era of high-throughput technologies, comparative transcriptomics studies in teleost fishes like the fugu and *Tetraodon* will extend our current understanding of vertebrate emrbyogenesis. However, the current genome assembly for fugu is highly fragmented (Jaillon et al., 2004), which makes the *Tetraodon* genome a better resource to study embryogenesis.

6.1.2 The role of Maternal to Zygotic transition during embryogenesis

After the commencement of fertilisation the embryo is subjected to accelerated cell divisions (Newport and Kirschner, 1982) followed by protraction of the cell cycle leading to the initiation of zygotic transcription and degradation of transcripts of maternal origin (Tadros and Lipshitz, 2009). The successive events, which lead to the elimination of maternally encoded products and the activation of zygotic transcription are collectively defined as Maternal to Zygotic Transition (MZT). A complex profile of transcriptional abundance is observed during the maternal to zygotic transition in various metazoans like *C.elegans* (Baugh et al., 2003), *Drosophila* (Arbeitman et al., 2002), *Ciona intestinalis* (Azumi et al., 2007), *Xenopus tropicalis* (Paranjpe et al., 2013), mouse (Hamatani et al., 2004) and human (Kocabas et al., 2006). Apart from coding genes, lncRNAs are also reported to be expressed in a stage specific fashion during MZT in *Xenopus tropicalis* implicating them to play an important role during embryogenesis (Paranjpe et al., 2013). In fact

lncRNAs are known to regulate various processes during vertebrate embryogenesis from microRNA-induced mRNA degradation to alteration of chromatin state (Pauli et al., 2011b). However little is known about the pattern of transcription and gene regulation during embryogeneis in teleost fishes, except for zebrafish (Aanes et al., 2011; Harvey et al., 2013; Mathavan et al., 2005) and medaka (Kraeussling et al., 2011). Unraveling the molecular mechanisms underlying the embryogenesis in diverse teleost fishes, will provide a better understanding on the genetic controls governing the diversification of body forms.

6.1.2 The early developmental transcriptome of *Tetraodon nigroviridis*

Currently no study has been undertaken to map the transcriptional repertoire of *Tetraodon* during early embryonic development. The primary causal factor is the lack of ready availability of *Tetraodon* eggs and embryos for research purposes (Watson et al., 2009). However, Watson *et al* reported a technique for successful breeding and spawning of the fish in laboratory environment. Further, Craig Watson provided his expertise to breed and spawn *Tetraodon* in the laboratory of my external supervisor Dr. Ferenc Müller. The small number of embryos collected at various developing stages led to the extraction of total RNA without replicates for sequencing. I have assembled and annotated the early developmental transcriptome in *Tetraodon* to analyse the transcriptional dynamics of both coding and long-non-coding genes. The aim of the experiment is to identify the coding and the long non-coding transcripts of maternal and zygotic origin involved in early development. Further, I wanted to compare the expression abundance of the

genes reported to be involved in MZT in zebrafish with their orthologs in *Tetraodon*. Finally, the assembly and annotation of the transcriptome is proposed as a resource to improve the structure and annotation of the existing gene models.

6.2 Materials and methods

6.2.1 RNA extraction and sequencing

Breeding, extraction of RNA and sequencing of *Tetraodon* eggs and embryos was performed in the laboratory of my external supervisor Dr. Ferenc Müller (A Zaucker, T Bodur, J Gehrig, Y Hadzhiev, F Loosli, H Roest Crollius, C Watson, F Müller, in preparation). Total RNA was extracted with Trizol (Invitrogen) according to the manufacturer's protocol from eggs and whole embryo at 30% epiboly (30 epi) and whole embryo at 24 hours post fertilisation (24 hpf). The RNA samples were treated with 2U Dnase I (Qiagen) per µg RNA sample at 37°C for 10 minutes. Digested samples were then treated with 20 mg/mL proteinase K (Sigma Aldrich) at 37°C for 45 minutes. The quality and quantity of total RNA were assessed with the Bioanalyzer 2100 (Agilent) and no sign of degradation was detected (RIN > 9.0). Sequencing libraries were generated from total RNA samples following the Truseq RNA protocol (Illumina). Single end reads (1 x 50 nucleotides) were obtained from 3 lanes on a Hiseq1000 using SBS v3 kits (Illumina). Cluster detection and base calling were performed using RTAv1.13 (Illumina). Quality of reads was assessed with CASAVA v1.9. Sequencing reads with a mean Phred score > 37 were further considered for mapping and assembly.

6.2.2 Quality filtering, mapping and assembly of sequenced reads

The raw sequencing reads from eggs, 30 epi and 24 hpf were processed with the Trimmomatic program (Lohse et al., 2012) to trim low quality bases, filter reads with low quality and filter reads smaller than 36 bases after trimming (parameters: ILLUMINACLIP::2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 HEADCROP:5). The raw reads were mapped on the *Tetraodon* genome (tetNig2) using the Tophat2 software (v2.0.8b) (Kim et al., 2013a) (parameters: --GTF --library-type fr-unstranded --segment-length 21 segment-mismatches 1 --raw-juncs –prefilter-multihits). A reference gene model file in the Gene Transfer Format (GTF) was used while mapping the reads. The reference GTF file comprised of pooled genomic features from Ensembl genes (Flicek et al., 2012b), transmap mRNA and transmap refgene tracks of the UCSC genome browser for *Tetraodon* (Meyer et al., 2012). The Cufflinks program (v2.1.1) (Trapnell et al., 2010) was used to assemble the reads mapped by using the chosen mapping strategy (parameters: --frag-bias-correct --library-type fr-unstranded --upper-quartile-norm --no-effective-length-correction). The transcript models generated by Cufflinks for the egg, 30 epi and 24 hpf mappings were merged together by the Cuffcompare utility from the Cufflinks software package (-V -R -r -s -C). All assembled transcripts longer than 200 bases were considered further.

6.2.3 Annotation of assembled transcripts

Reference file of gene models in GTF format was obtained from Ensembl (ftp://ftp.Ensembl.org/pub/release-73/gtf/Tetraodon_nigroviridis/). The reference GTF file from Ensembl was converted in refFlat format (http://genome.ucsc.edu/goldenPath/gbdDescriptionsOld.html#RefFlat) using the UCSC utility gtfToGenePred (http://hgdownload.cse.ucsc.edu/admin/exe/) and a custom script in the Perl language. The reference file in refFlat format was compared with the mapped reads from eggs, 30 epi and 24 hpf to extract the percentage of reads mapping to different genomic features using the CollectRnaSeqMetrics.jar utility from the picard tools software package v1.88 (http://picard.sourceforge.net/). The Annocript pipeline was used to annotate the assembled transcript sequences. All transcripts assigned an identifier from SwissProt or Uniref90 or Conserved Domain Database (CDD) during the BLAST sequence homology comparison are predicted to be coding. All transcripts which do not have an annotation from SwissProt, Uniref90, CDD and Rfam and contain an ORF smaller than 100 amino acids are considered as Potential Long Non-Coding sequences (PLoNCs). The non-coding potential (NCP) of all PLoNCs was predicted by Annocript using the Portrait software (Arrial et al., 2009). A score greater than the mean NCP score of all PLoNCs (0.76) was used to predict the final lncRNA set from the PLoNCs. The overlap of coordinates of the predicted lncRNAs with the predicted coding transcripts and coding genes from Ensembl (v74) was checked with the intersectBed program from the BEDTools package. All lncRNA transcripts not overlapping a coding loci are classified as long intergenic non-coding RNAs (lincRNAs). The assembled transcripts were mapped to the reference Ensembl GTF file (v73) using Cuffcompare (-V -R -r -s -C).

The results of the mapping was used to infer the number of assembled transcripts which are predicted to be novel isoforms of existing gene models.

6.2.4 Generation of Circos map

The *Tetraodon* genome was divided into 1 megabase bins using the windowMaker utility from BEDTools software package v 2.17 (Quinlan and Hall, 2010). The RPKM expression value for each assembled loci was obtained by calculating the mean RPKM for all transcripts falling in that particular loci. The intersectBED utility from BEDTools was used to find the intersection of assembled coding and long non-coding loci with the genome wide 1 MB bins. The expression intensity of coding and lncRNA transcripts in each genomic bin was considered as the mean RPKM of all coding and lncRNA loci falling inside a genomic bin. MultiZ (Blanchette et al., 2004) alignment of eight vertebrate genomes with the zebrafish as reference (other species: human, mouse, medaka, stickleback, fugu, *Tetraodon*, *Xenopus tropicalis*) was downloaded in the Multiple Alignment Format (MAF) from the UCSC database (http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/multiz8way/multiz8way.maf.gz). Genome wide alignment scores for *Tetraodon* were obtained in BED format from the MultiZ alignment using the mafSpeciesSubset utility from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/) along with a custom perl script. The intersectBed utility was used to find the intersection between the genome wide 1 MB bins in *Tetraodon* and the genome wide aligned regions from the MultiZ alignments. The mean conservation score of all conserved regions falling within a genomic bin was assigned as the conservation score for that particular genomic bin. The frequency

of coding and long non-coding loci within each genomic bin was calculated using the coverageBed utility from the BEDTools software package. The conservation scores and expression intensity for coding and lncRNA loci in each genomic bin was formatted according to the instructions given in the manual of the Circos software v0.64 (Krzywinski et al., 2009) using custom scripts in perl language. The Circos perl script was run on the formatted data to generate the circular image for the complete transcriptome experiment.

6.2.5 Detection of sequence conservation and visualisation in the genome browser

The intersectBed utility from BEDTools package was used to find the intersection of the exonic coordinates of the assembled transcripts with the MultiZ alignments obtained in the previous section. The product of the conservation score of a conserved region falling within a exon, with the fraction of overlap is taken as the conservation score for the exon. The sum of scores of all exons of a transcript is taken as the conservation score for the transcript. The output files from Tophat2 in BAM format were converted to BigWig format using the genomeCoverageBed binary from the BEDTools package (v2.17) (Quinlan and Hall, 2010) and the bedGraphToBigWig utility from the UCSC database (http://hgdownload.cse.ucsc.edu/admin/exe/). The visualisation of the RNAseq peaks and transcript models was carried out in the Integrative Genomics Viewer (v2.2.7) (Thorvaldsdóttir et al., 2012).

6.2.6 Differential expression analysis of the assembled transcripts

The raw read counts for all exons of assembled transcripts were obtained with the multiBamCov utility from BEDTools software package (-split). The sum of read count for all exons of a given transcript was considered as the raw count of the transcript. The bioconductor edgeR package (Robinson et al., 2010) was used to calculate the differential expression of transcripts across the developmental stages. This package measures the significance of the variation in expression levels using the dispersion of the expression levels among sample replicates. In the absence of replicates the software can infer the dispersion value using the fluctuations in the expression levels of selected house-keeping genes among the different samples. Given that the analyzed dataset was only composed by single samples for each stage (i.e. no replicates) and there is no information about housekeeping genes in *Tetraodon*, I used the following strategy to infer an acceptable dispersion value. All the transcripts showing less than 1 read per million (mapped) in the sum of the experiments were discarded to filter lowly expressed or background-biased transcripts. For all the remaining transcripts I calculated the standard deviation among the expression levels. Transcripts were then sorted in descending order on the standard deviation values. Next, I filtered out all the genes that did not get any match against the UniRef database in the annotation step. Then, based on the standard deviations and the annotations, I selected the transcripts showing the lowest variations until I was able to collect 100 different genes (based on the annotations). These 100 genes, corresponding to 305 transcripts, were considered bona-fide house-keeping, and the dispersion value was calculated using them. In

addition, in order to improve the stringency of the selection and to further reduce the number of potential false positives in the differential expression analysis, I multiplied the obtained dispersion value by 10. The calculated dispersion value (0.5) was used for the differential expression analysis in edgeR. Transcripts with more than 0.5 reads per million mapped reads in at least one sample were retained for the analysis. The following comparison were executed using the exactTest function with default parameters: eggs vs 30 epiboly, eggs vs 24 hours post fertilisation, 30 epiboly vs 24 hours post fertilisation. According to the comparison, significantly up/downregulated transcripts were selected, considering all the transcripts with a FDR value smaller than 0.05 and a linear fold change of at least 2 folds. The maternal specific list of transcripts was prepared selecting only those transcripts resulting significantly up-regulated in the eggs in both the comparisons involving the eggs sample. The embryonic list of transcripts was prepared selecting only those transcripts resulting significantly down-regulated in the eggs in both comparisons involving the eggs samples.

6.2.7 Identification of microsynteny, prediction of sequence conservation and gene ontology enrichment

The SynLinc pipeline was used to predict putative microsyntenic lincRNAs between *Tetraodon* and zebrafish, considering only immediate flanking coding genes for each lincRNA. The gene ontology (The Gene Ontology Consortium, 2012) enrichment analysis was performed on the GO mapping done by the Annocript pipeline using a custom R script exploiting the Fisher exact text and p-value FDR

correction to select significantly enriched GO classes (minimum representatives for a GO class: 5; FDR <= 0.05).

6.2.8 Comparison of expression abundance between maternal and zygotic transcripts in zebrafish with their *Tetraodon* orthologs

The expression of zebrafish genes in FPKM during early development and the lists of genes reported to be maternal, zygotic and maternal/zygotic were obtained from a recent published report (Harvey et al., 2013). The expression (FPKM) of the assembled transcripts in *Tetraodon* was calculated by Cuffdiff program from the Cufflinks software package (--frag-bias-correct --multi-read-correct --no-effective-length-correction --upper-quartile-norm --max-frag-multihits 20). The mean FPKM of all transcripts mapped to an Ensembl gene was considered to be the expression for that gene. Orthologous protein coding genes between zebrafish and *Tetraodon* was obtained with the Bioconductor (Gentleman et al., 2004) biomaRt (Durinck et al., 2005) package.

6.3 Results and Discussion

6.3.1 Mapping, assembly and annotation of the early developmental transcriptome of *Tetraodon*

More than 200 million short reads were generated by sequencing of the RNA samples from whole embryo during three developmental stages, eggs, 30% epiboly (30 epi) and 24 hour post-fertilisation (24 hpf) in *Tetraodon* (Eggs: 229,741,278; 30

epi: 248,151,367; 24 hpf: 239,853,692). In excess of 90% of reads from all stages passed the quality filtering tests (Eggs: 215,219,078; 30 epi: 226,292,377; 24 hpf: 224,089,479). I wanted to compare the aligned reads to various genomic features defined in *Tetraodon* by the Ensembl gene build (v73). Hence, I compared the positions of the short reads aligned to the genome with the existing gene models from Ensembl (**Figure 6.1**). Majority of the reads overlap coding exonic regions while a smaller fraction fall in intergenic locations. This confirms that a significant percentage of transcription occurs in the protein-coding loci during embryogenesis in *Tetraodon*. However more than 10% of mapped reads fall under intergenic loci suggesting the active participation of non-coding elements in processes regulating early development of the embryo. Past reports indicate that the cellular miRNA machinery governs the molecular pathways involved in degradation of maternal mRNAs, and controls the spatial and temporal expression of embryonic mRNAs (Giraldez, 2010; Svoboda and Flemr, 2010). However the RNA extraction for *Tetraodon* was not optimised to enrich for small RNA fraction, hence the intergenic transcription probably represents other classes of non-coding RNAs, especially lincRNAs.

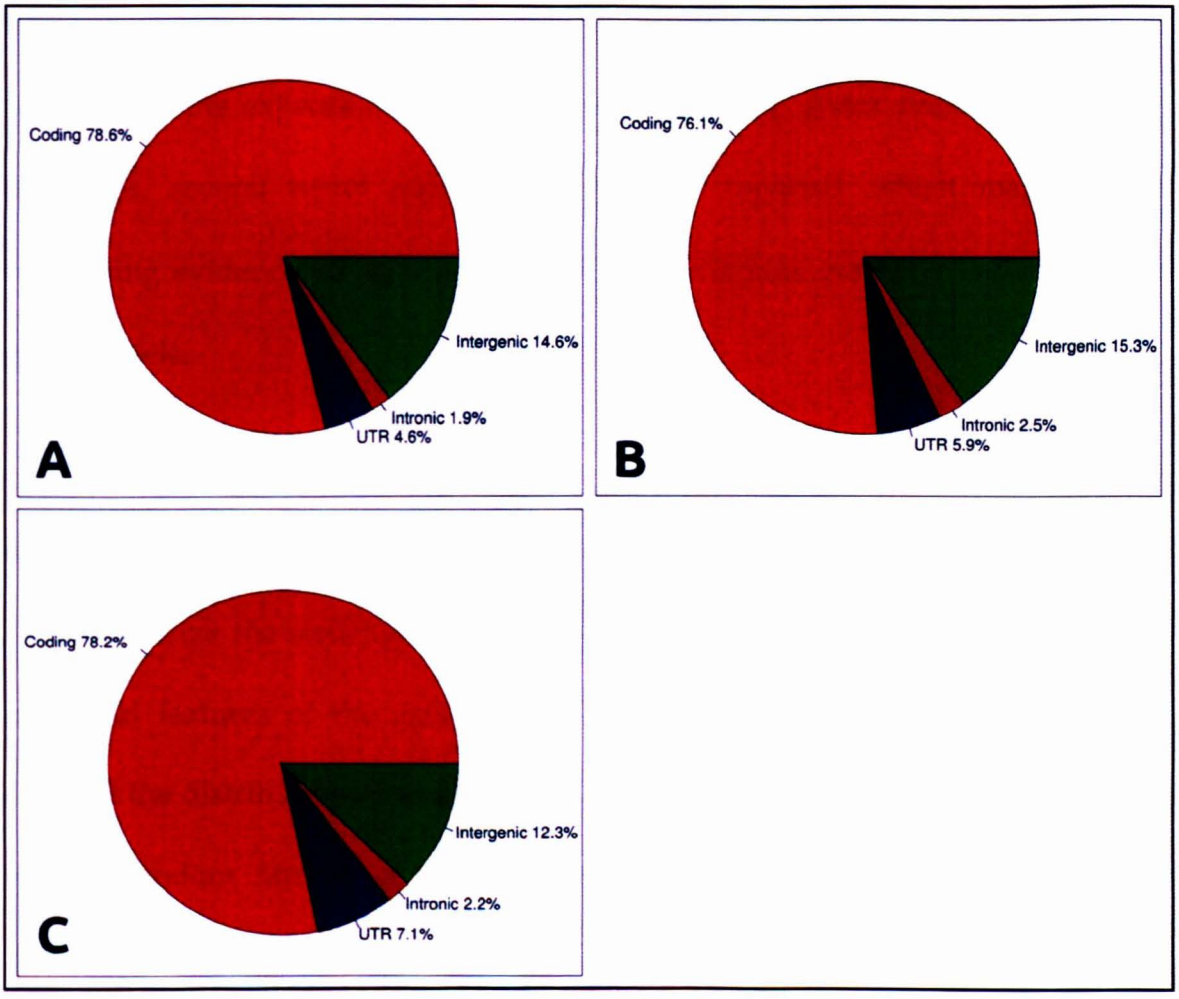


Figure 6.1 Percentage of aligned short reads overlapping genomic features predicted for the *Tetraodon* genome by Ensembl (v74) in **A)** Eggs **B)** 30% Epiboly **C)** 24 hours post-fertilisation.

The mapped reads resulted in the assembly of 61,033 transcripts from 23,838 loci on the *Tetraodon* genome. I predicted 53,543 coding transcripts from 19,414 loci and 4026 long non-coding transcripts from 3508 loci. A subset of lncRNA transcripts (2994 transcripts from 2663 loci) fall in intergenic regions, hence were classified as long intergenic non-coding RNAs (lincRNAs). A majority of the existing Ensembl coding gene models (86%) are represented by the assembled coding transcripts while 3036 loci comprising of 5093 coding transcripts are not present in the

Ensembl annotation. This shows that, while the transcript assembly is able to account for the expression of the majority of coding genes present in genome databases, several novel coding loci are also reported which may provide supporting evidence for annotation of genomic regions currently lacking proper gene models.

6.3.2 Genomic structure and conservation of the early developmental transcripts

To characterize the structure of the generated transcript models, I compared the structural features of the assembled coding and lncRNA transcripts. Hence, I checked the distribution of length, number of exons and exon size of the lncRNAs with a random sample of coding transcripts (**Figure 6.2**). The lncRNAs are observed to be smaller in length with fewer but longer exons in comparison to the coding transcripts. Majority of the predicted lncRNAs (64%) are not spliced (2593 transcripts). I observed a similarity of the structural properties with the lncRNA transcripts expressed in the zebrafish islet cell (**Figure 5.8**). As discussed in the previous Chapter 5, long non-coding RNAs tend to be less spliced than coding genes, hence the observation falls in agreement with a general lncRNA property.

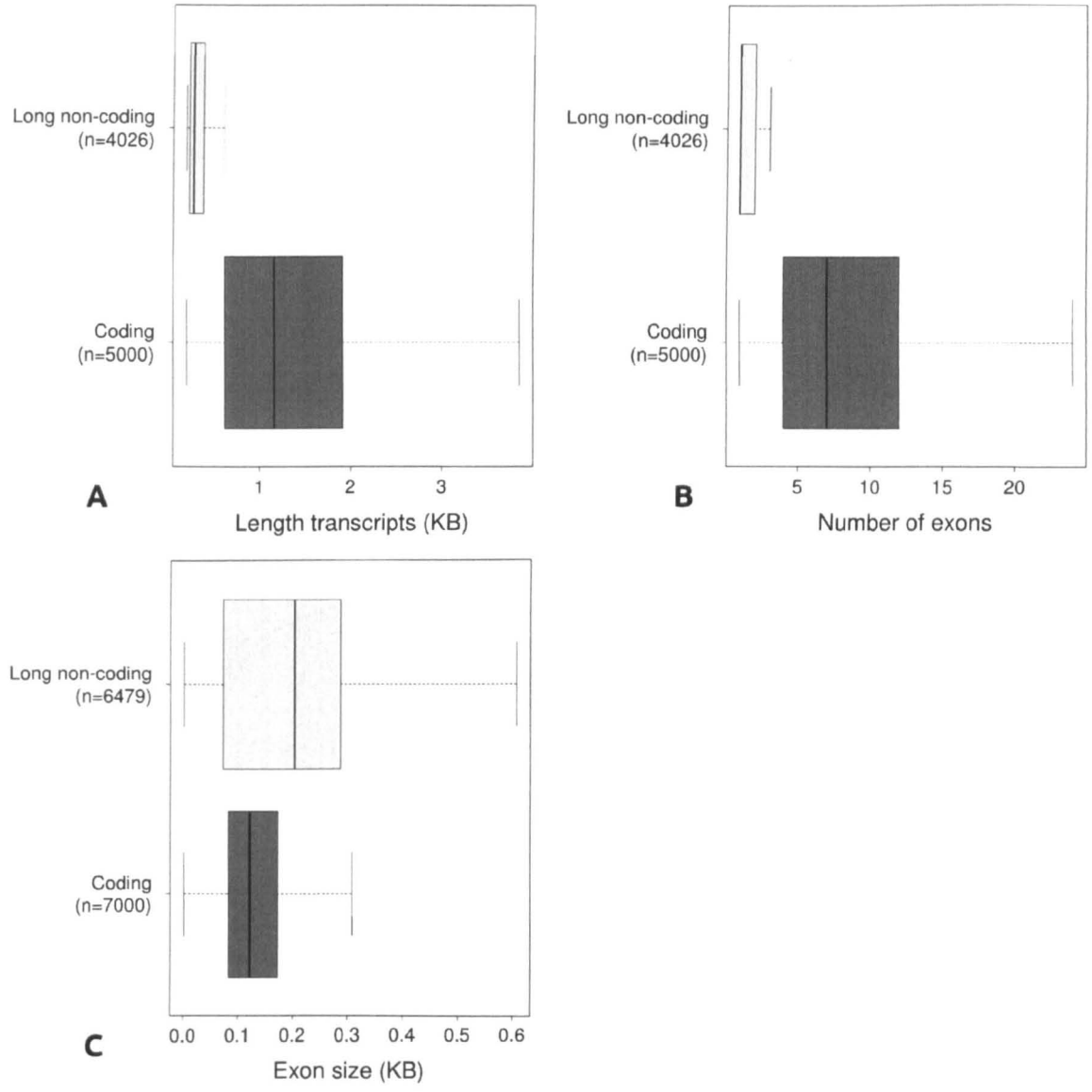


Figure 6.2 Structural features of coding and long non-coding transcripts **A)** Length of transcripts **B)** Number of exons **C)** Size of exons.

Then, I compared the genomic location of the coding, lncRNA and lincRNA transcripts with whole genome alignments of 8 vertebrate species to extract the sequence conservation metric for each feature. In agreement with previous reports in other species (Cabili et al., 2011; Guttman et al., 2010; Pauli et al., 2011a) as well as the zebrafish islet cell lncRNAs **(Figure 5.9)** the *Tetraodon* lncRNAs show a lower level of sequence conservation than coding transcripts but are marginally better conserved than random intergenic regions (**Figure 6.3**). However, contrary to the

observation in zebrafish islet cell lncRNAs, I did not find a significant difference in conservation between lncRNA and lincRNA transcripts in *Tetraodon*. This may be explained by the fact that the lincRNAs are a subset of lncRNAs. Since the majority of *Tetraodon* lncRNAs belong to lincRNA class (74%) in comparison to zebrafish islet cell transcriptome (22%) there is no marked difference in their conservation levels.

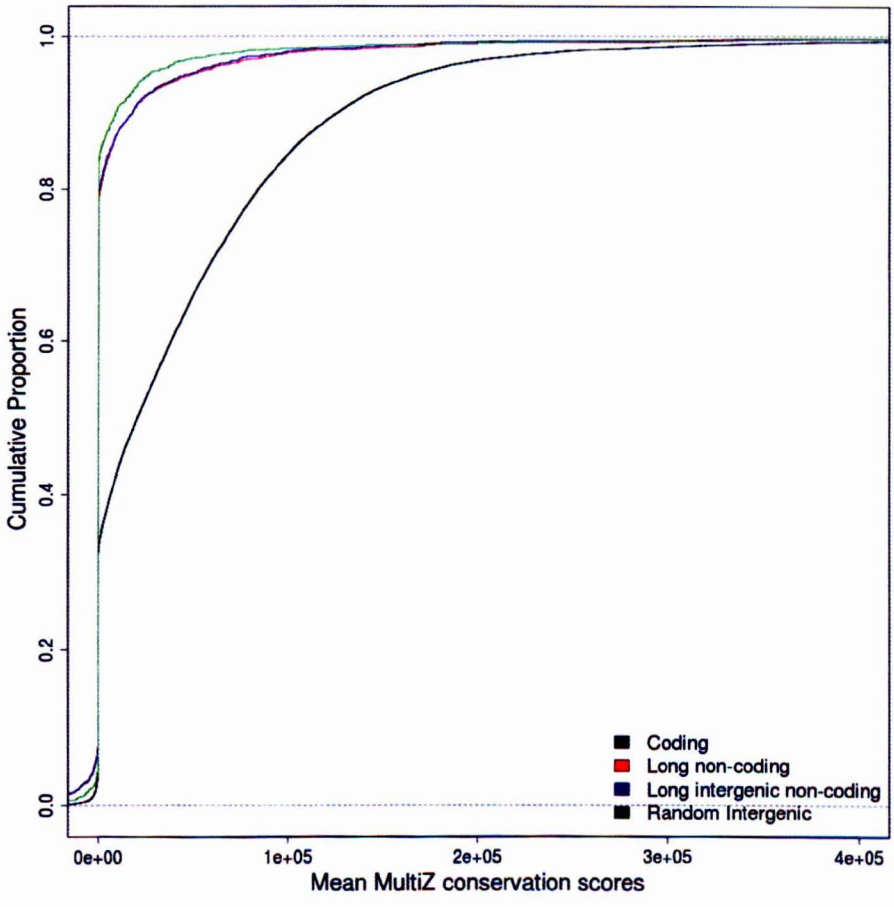


Figure 6.3 Mean MultiZ 8 way whole genome alignment scores of *Tetraodon* for coding, long non-coding and long intergenic non-coding transcripts. The alignment scores are calculated by aligning the zebrafish genome with the human, mouse, *X. tropicalis*, *Tetraodon*, medaka, fugu and stickleback genomes. The x-axis represents the mean MultiZ8way alignment scores and the y-axis represents the cumulative proportions of the transcripts.

6.3.3 Inspection of expression dynamics of coding and long non-coding loci during *Tetraodon* development

The RNAseq experiment was designed to obtain a temporal snapshot of the coding and the long non-coding transcripts during early development in *Tetraodon*. I decided to generate a single composite image which may illustrate the complete experiment and help in selecting genomic regions which show an interplay of expression between coding transcripts and lncRNAs. In this context I used the Circos (Krzywinski et al., 2009) software to generate an image depicting the average expression of coding and long non-coding transcripts across 1 megabase bins of the *Tetraodon* genome in the 3 developmental stages (**Figure 6.4 A**). The circos image provides a summary of the experiment and, at a glance, the coding loci show a more homogenous pattern of expression across the whole genome, while the lncRNA loci are expressed in sparse pockets. The aim was to identify genomic loci where the expression dynamics of coding transcripts and lncRNAs suggest a coherent regulatory mechanism to promote organism development. It is important to note though, that the intensity of colour in a genomic bin is based on average expression (RPKM) of all features in that bin (coding or lncRNA), hence a few highly expressed genes can result in the Circos image showing a higher expression abundance for the complete genomic bin. I inspected several genomic bins, where both coding and lncRNA loci are dynamically expressed during development. Among them, a very interesting genomic bin pair is on chromosome 21. Here a first bin (chr21:3,000,000-4,000,000) shows a decrease in average expression intensity of the coding loci (from egg to 24 hpf), which is

complemented by high average expression of the lncRNA loci in both eggs and 30 epi and low expression in 24 hpf (**Figure 6.4 B**). Conversely, the preceding coding genomic bin (chr21:2,000,000-3,000,000) shows a rise in expression of coding loci from egg to 24 hpf but no relevant lncRNA transcription. The genomic region at the bounduary between these 2 bins contains the *Tetraodon Hoxa* gene cluster (chr21:2,996,402-3,093,266), which I decided to investigate further.

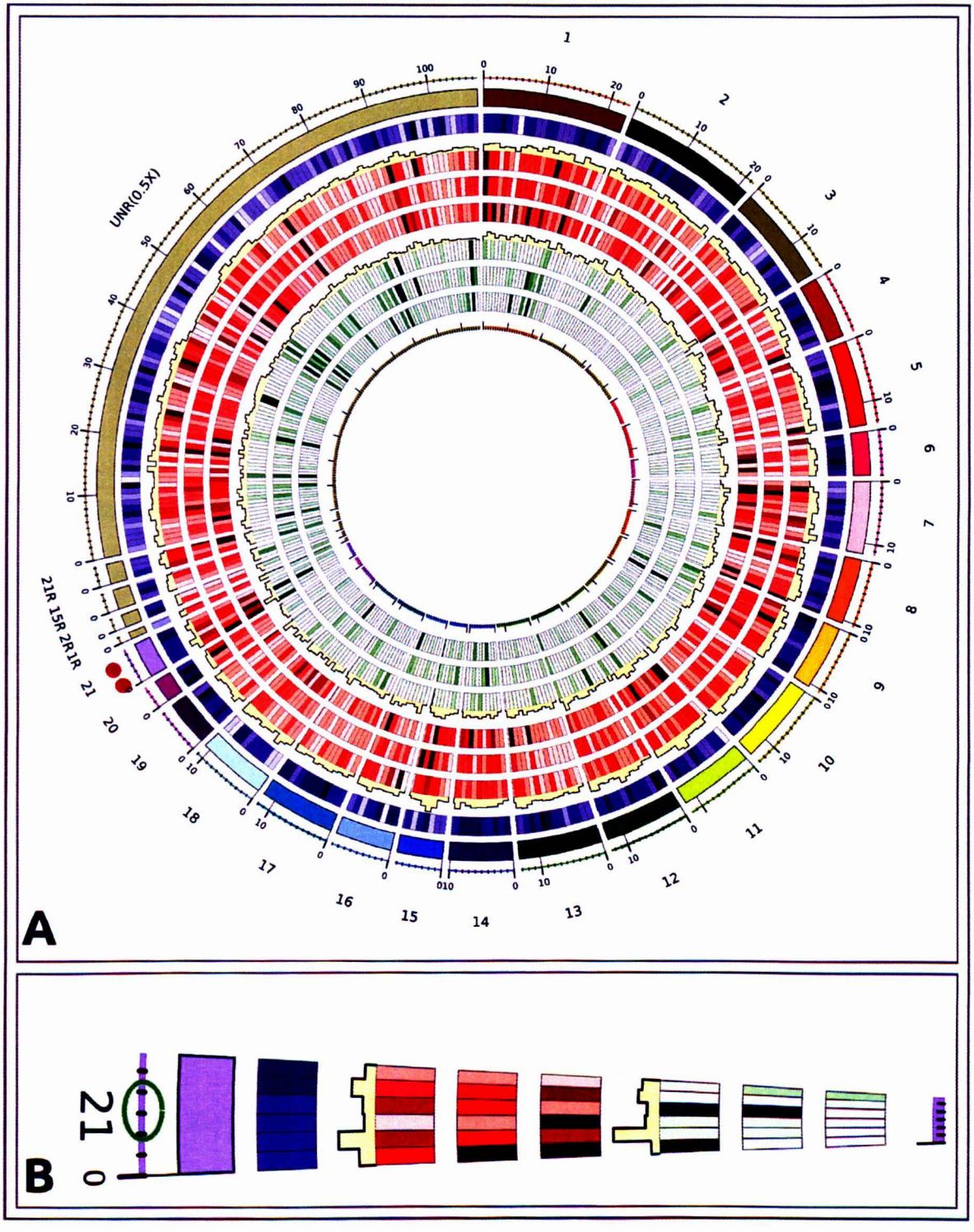


Figure 6.4 Circos image depicting the average expression of coding transcripts and lncRNAs in 1 MB bins across three developmental stages in the *Tetraodon* genome. A) The outermost circle represents the *Tetraodon* chromosomes, divided into 1 MB bins. The next circle (purple) shows the average sequence conservation score of each bin across eight vertebrate species. The next six circles show the mean expression

(RPKM) of the sum of all coding transcripts (red) and lncRNAs (green) falling within each chromosomal bin in eggs (outer), 30 epi (middle) and 24 hpf (inner). The histogram above the circle showing mean expression of eggs represents the frequency of each transcripts class in each bin **B)** The magnified view of chromosome 21 which contains the *Hoxa* gene cluster in between the 3rd and 4th bin (marked by green circle). Note: The intensity of color reflects the level of expression and conservation (dark to light → high to low). The red dots mark the candidate genomic bin showing coordinated expression of coding and lncRNA loci.

The vertebrate *Hoxa* cluster genes are involved in the development of craniofacial skeleton, vertebrate limb-bud and pectoral and caudal fin in jawed fishes (Gardiner et al., 1995; Géraudie and Borday Birraux, 2003; Minoux et al., 2009). The *Tetraodon Hoxa* genes show a coordinated expression at 24 hpf and few genes are also expressed maternally and at 30% epiboly. The role of lncRNAs during early embryogenesis and regulation of *HoxA* genes is well known in mammals. The lncRNA *Hoxa* Transcript at the distal Tip (*HOTTIP*) is reported to lie at the posterior end (upstream of *Hoxa13a*) of the human *Hoxa* cluster and activates the transcription of several *Hoxa* genes by altering the chromatin state of their genomic loci (Wang et al., 2011). Another lncRNA the *Hox* Antisense Intergenic RNA Myeloid 1 (*HOTAIRM1*) is expressed between the human *Hoxa1* and *Hoxa2* genes and is implicated in myelopoiesis by regulating the expression of *Hoxa1* and *Hoxa4* genes (Zhang et al., 2009). I did not find location specific homologs of the *HOTTIP* and *HOTAIRM* lncRNAs in the *Tetraodon Hoxa* clusters. However, a lncRNA *linc_hoxa1a* is present at the anterior most end of the one *Hoxa* cluster (downstream to the *Hoxa1A* gene), while another lncRNA *lnc_hoxa7a* lies between the *Hoxa7* and

Hoxa5a genes (**Figure 6.5 A**). The second *Hoxa* cluster (**Figure 6.5 B**) does not contain any lncRNAs in its vicinity. The zygotic expression of the *lnc_hoxa1a* transcript is concurrent with the zygotic expression of *Hoxa1A* and *Hoxa2a* during 30 epi, while all the other members of the cluster remain untranscribed. Then, at 24 hpf, all the genes assembled and annotated in the cluster result expressed, while the *lnc_hoxa1a* appears silent. It is intriguing to speculate whether the *lnc_hoxa1a* is implicated in the activation of the transcription of the *Hoxa* cluster, but a higher number of developmental time points, earlier than 30 epi, is required to properly answer this question. Nevertheless, the *Hoxa* cluster locus demonstrates the co-expression of long non-coding genes with proximal coding genes, which are known to play fundamental roles in organism development.

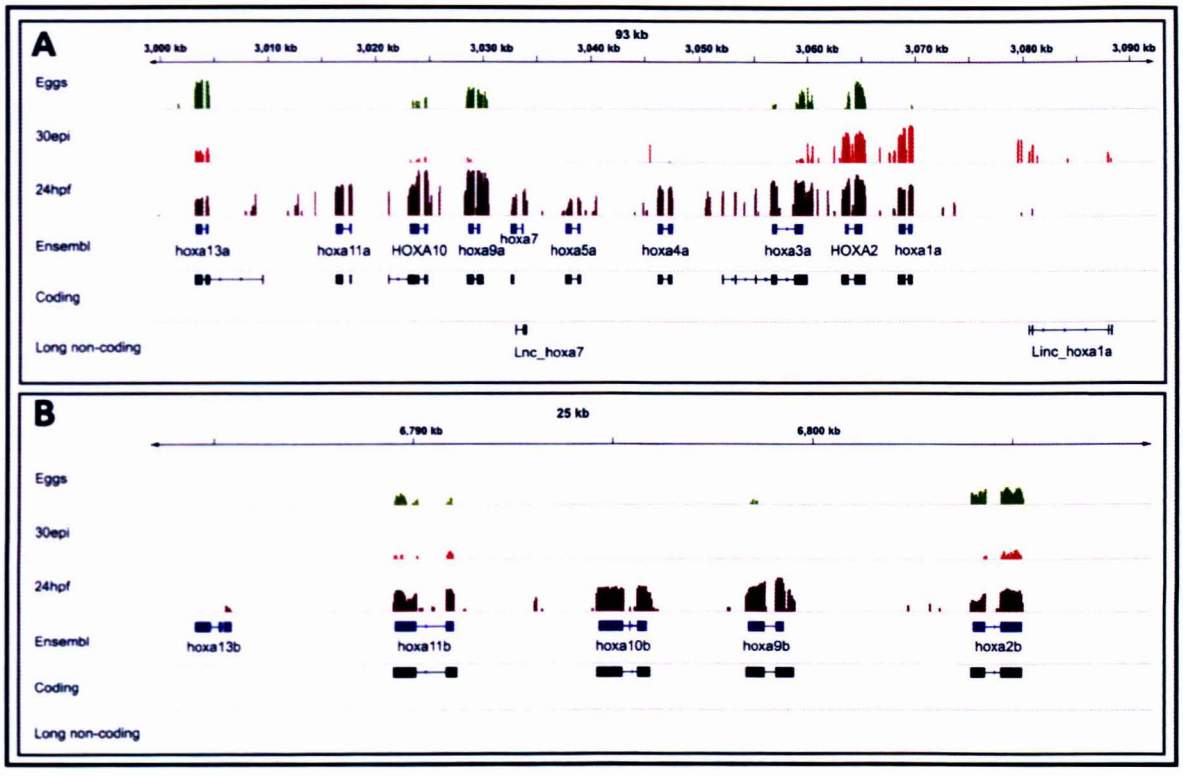


Figure 6.5 The *Hoxa* cluster of genes in the *Tetraodon* genome. **A)** The first *Hoxa* cluster (chr21:2,996,402-3,093,266) **B)** The second *Hoxa* cluster (chr8:6,782,441-6,811,548). The tracks Eggs, 30 epi, 24 hpf show the coverage depth of mapped reads on the genome during different developmental stages. The Ensembl track contains gene models defined by the Ensembl database. The Coding and Long non-coding tracks contain gene models assembled from the mapped reads across the three developmental stages.

6.3.4 Maternal and embryonic specific transcripts in *Tetraodon*

Based on a differential expression analysis of the assembled transcripts, I defined maternal and embryonic specific transcripts from the transcriptome assembly (see methods; log2FoldChange >= 2; FDR <= 0.05). I separated the maternal and embryonic list into coding and lncRNA transcripts based on the annotations by Annocript (**Table 6.1**). I found that the average expression of a lncRNA is higher at 30 epi in comparison to coding transcripts (two sample proportion test, p-value:

1.135e-15) (**Figure 6.6**). Long non-coding RNAs are reported to show a coordinated pattern of expression with coding genes implicated in early organism development (Dinger et al., 2008; Guttman et al., 2011). However, lncRNAs are also reported to be expressed in a very small time window during the early stages of embryogenesis in the zebrafish (Pauli et al., 2011a). Thus the higher expression level of the differentially expressed lncRNAs at 30 epi could result from their participation in a diverse set of regulatory programs important during early vertebrate embryogenesis.

Class	Coding transcripts	LncRNAs	LincRNAs
Differential in any stage	9533	401	301
Maternal	222	11	7
Embryonic	3625	195	152

Table 6.1 Number of transcripts showing differential expression.

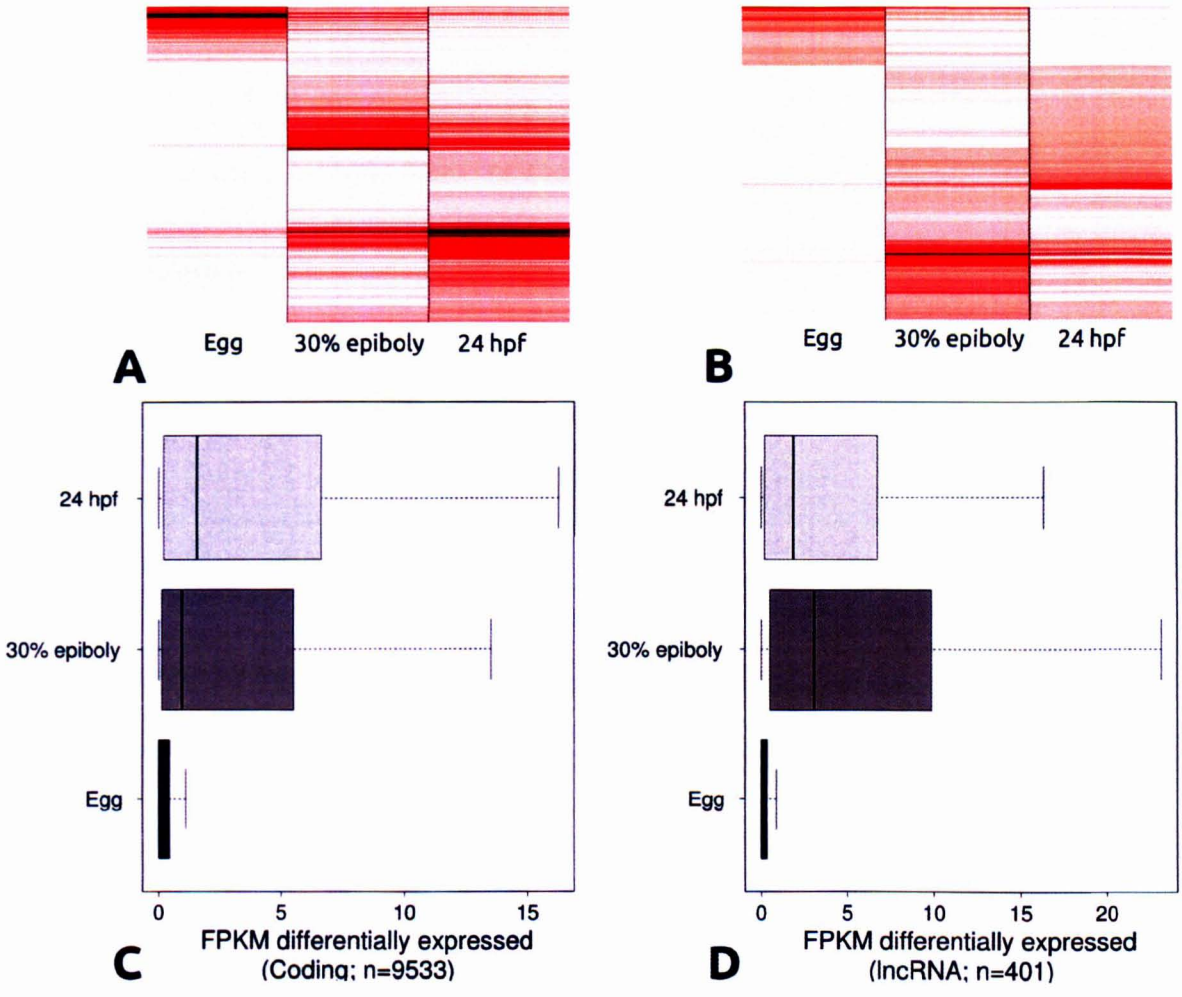


Figure 6.6 Expression dynamics of differentially expressed coding and long non-coding transcripts during early development in *Tetraodon*. **A)** Heat map of expression abundance (RPKM) for differentially expressed coding transcripts **B)** Heat map of expression abundance (RPKM) for differentially expressed lncRNA **C)** Distribution of expression abundance (RPKM) for differentially expressed coding transcripts in different developmental stages **D)** Distribution of expression abundance (RPKM) for differentially expressed lncRNAs in different developmental stages.

6.3.5 Gene ontology enrichment of maternal and zygotic coding transcripts in *Tetraodon*

To discern the classes of coding genes being differentially expressed during early development in *Tetraodon,* I performed a gene ontology enrichment analysis. I took

a two pronged approach where I performed the analysis separately for all coding genes and for coding genes present in the the vicinity of lincRNA transcripts (10KB upstream or downstream). The maternal differentially expressed coding genes show a significant enrichment for GO classes like *determination of dorsal identity, regulation of WNT signalling pathway, actin cytoskeleton organisation* and *mitosis* (**Figure 6.7**).

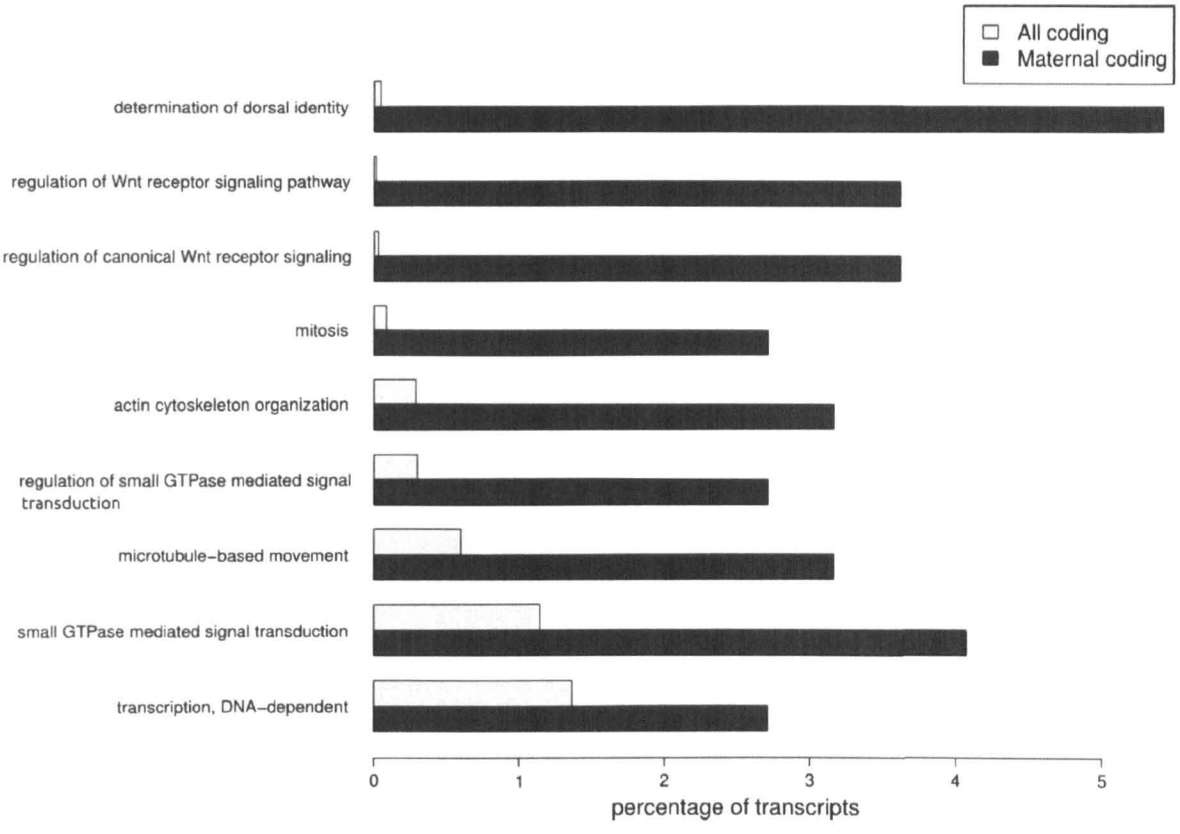


Figure 6.7 Gene ontology enrichment analysis for differentially expressed maternal coding genes in *Tetraodon*. The x-axis represents the percentage of transcripts which are defined by a particular GO biological process and the y-axis represents the significantly enriched GO classes.

Past studies have confirmed the role of maternally introduced transcripts during specification of the embryonic axes in *Drosophila* (Grünert and St Johnston, 1996), *Xenopus laevis* (Mowry and Cote, 1999) and zebrafish (Pelegri, 2003). Particularly in

anamniotes, a microtubule-dependent operation displaces maternal molecular factors from the vegetal pole towards the zone of future dorsal side of the embryo, thus disrupting the initial radial symmetry of the zygote (Weaver and Kimelman, 2004). This act of axis determination in general is dependent on maternally originating Wingless-Type MMTV Integration Site Family, Member 1/5a (*WNT11/5a)* complexes, under regulation of the *WNT* antagonist Dickkopf *WNT* signaling pathway inhibitor 1 (*DKK-1*) (Cha et al., 2009). A recent reports mentions the role of a maternal canonical *WNT* and not the *WNT/5a* complex specially in the case of dorsal axis determination (Lu et al., 2011). Hence the enriched GO classes are in agreement with past reports of important functions partaken by transcripts of maternal origin.

The process of *somitogenesis* is predicted to be the enriched GO class with the maximum representatives of embryonic specific transcripts (**Figure 6.8**). During embryonic development, *somitogenesis* results in formation of bilaterally paired mesoderm tissue along the anterior-posterior axis of the developing embryo. The somites are differentiated into muscle, cartilage, endothelial cells, and dermis. The signaling cascade of the Bone Morphogenetic Protein 4 (*BMP4*) protein is an important component regulating the differentiation of somites into muscle lineage (Tajbakhsh and Cossu, 1997). The *regulation of BMP signaling pathway* is predicted to be one of the significantly enriched GO term in the embryonic specific transcripts. While enriched GO terms like *calcium-dependent cell-cell adhesion, organ morphogeneis* and *regulation of transcription* are commonly associated with the

developing embryo it is unusual for the embryonic transcripts to be enriched for *RNA-dependent DNA replication*.

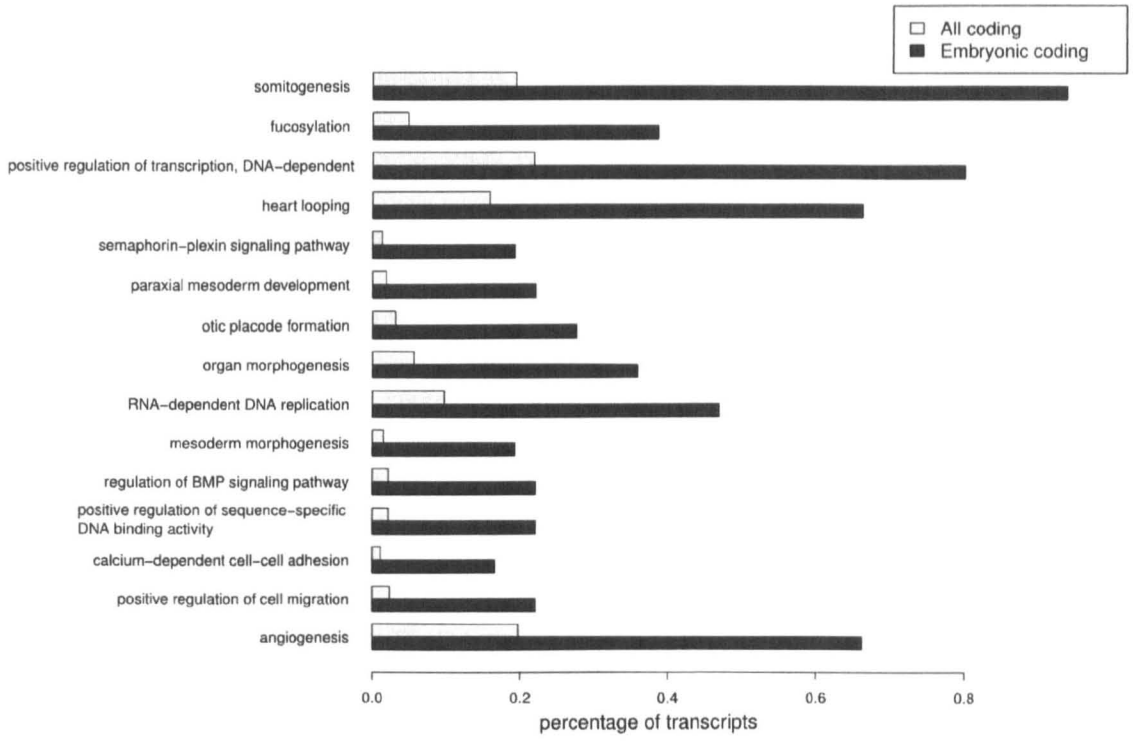


Figure 6.8 Gene ontology enrichment analysis for differentially expressed embryonic coding genes in *Tetraodon*. The x-axis represents the percentage of transcripts which are defined by a particular GO biological process and the y-axis represents the significantly enriched GO classes.

6.3.6 Gene ontology enrichment of maternal and zygotic coding transcripts flanking lincRNAs in *Tetraodon*

To throw further insights into the possible role of non-coding RNAs during development I compared the proximal coding genes of all lincRNAs (10KB upstream and downstream) against the proximal coding genes of lincRNA predicted to be differentially expressed in both 30 epi and 24 hpf (embryonic lincRNAs) (**Figure 6.9**). I chose only the intergenic lncRNAs (lincRNA) for my

analysis since it reduces the probability of a GO term to be associated with a lncRNA due to an overlapping coding gene. The number of lincRNAs predicted to be differentially expressed and of maternal origin were not large enough (7 transcripts) to perform a statistical analysis on the GO terms of proximal coding transcripts. The prominent GO terms significantly enriched to represent coding transcripts neighboring differentially expressed embryonic lincRNAs include *protein glycosylation, embryonic viscerocranium morphogenesis* and *nucleosome assembly*. The attachment of glycans (large carbohydrate molecules) to proteins or other organic molecules with the aid of enzymatic action is known as glycosylation which helps in intra-cellular transport, post-translational modification of proteins and metabolic homeostasis. The early developmental transcriptome in zebrafish contains a diverse population of glycans suggesting a complex glycosylation pattern during embryogenesis (Chang et al., 2009; Guérardel et al., 2006). In fact, proteins important during embryo development like the *WNT* require glycosylation process to induce modifications in structure to perform their specific function (Ke et al., 2013). There might be an involvement of lincRNAs in the recruitment of factors which aid in the cellular glycosylation although no direct experimental evidence exists in favor of the argument.

However a well studied aspect of lncRNA biology is the effect of their regulation over genes playing an important role in nervous system development (Qureshi and Mehler, 2012). The embryonic lincRNAs are enriched to lie near coding genes involved in *embryonic viscerocranium morphogenesis* a process which defines the

vertebrate craniofacial skeleton. Interestingly a dynamic expression pattern of glycoconjugates was reported to correlate with morphological modifications in the rat fetal viscerocranium indicating the glycosylation of proteins involved in establishment of the fetal brain structure (Zschäbitz et al., 1999). The neural-crest cells are an important factor governing the formation of the vertebrate craniofacial skeleton (Kague et al., 2012). The neural crest cells are a transient cell population in the vertebrate embryo which carry the potency to differentiate further into major cell types like melanocytes, craniofacial cartilage/bone, smooth muscle and neurons (Huang and Saint-Jeannet, 2004). It is worth noting that the GO term *neural crest cell fate specification* is also predicted to be enriched in my analysis further insinuating the role of lincRNAs in pattering of brain tissue and skeleton during early development of *Tetraodon.* Another aspect of lincRNA functioning is highlighted by the enrichment of the GO term *nucleosome assembly.* Intergenic non-coding RNA are reported to maintain a repressed chromatin state by directing increased nucleosome occupancy in their genomic neighborhood in yeast (Hainer et al., 2011). A recent study shows the Hepatic Nuclear Factor 1A antisense 1 (*HNF1A-AS1*) lncRNA to be involved in regulation of genes important for assembly of chromatin and the nucleosome, thus indirectly modulating the cell cycle progression (Yang et al., 2013b). In principle lincRNAs are projected as an important cog of the vertebrate developmental programming due to their ability of acting as a host scaffolding molecule for protein complexes which can specify the pattern of histone modifications in target genes to achieve a specific expression pattern aiding cellular development and differentiation (Blelloch and Gutkind,

2013; Lee, 2012). Hence the current analysis shows the differentially expressed embryonic lincRNAs to be associated with coding genes principally involved in specification of the vertebrate brain, body patterning and regulation of chromatin state.

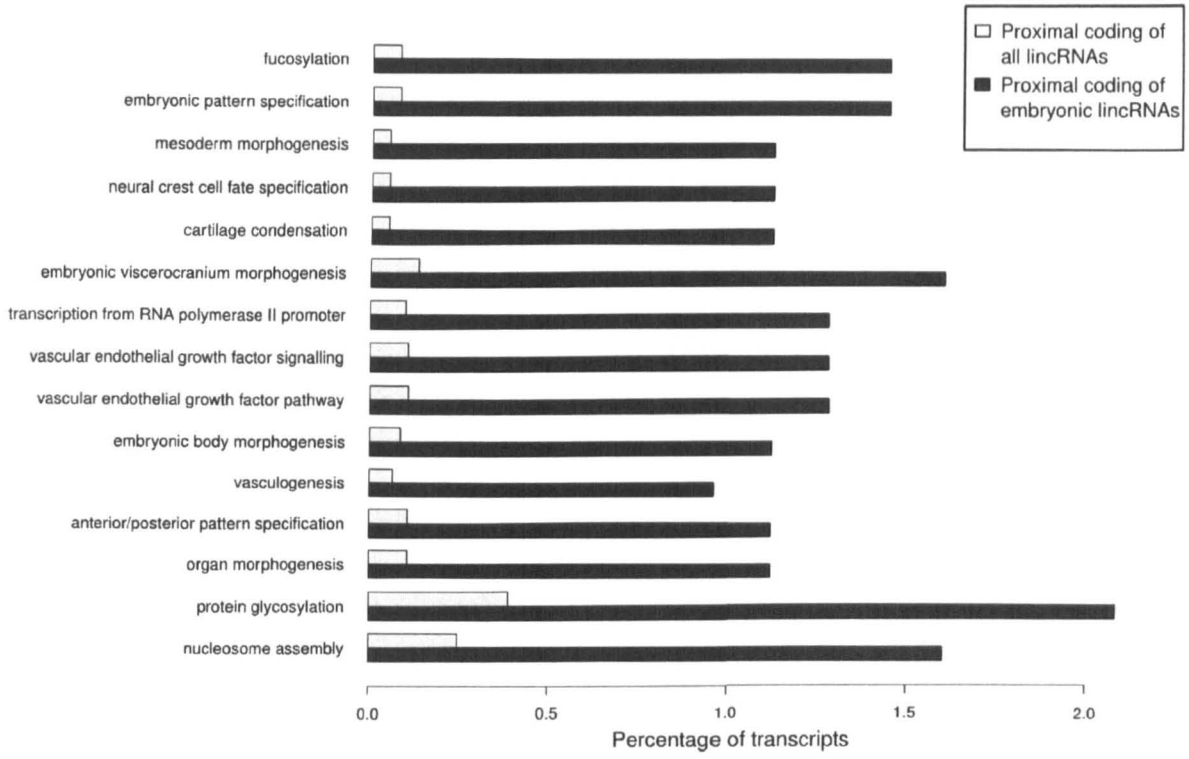


Figure 6.9 Gene ontology enrichment analysis for coding genes lying proximal to differentially expressed embryonic lincRNAs in *Tetraodon*. The x-axis represents the percentage of transcripts which are defined by a particular GO biological process and the y-axis represents the significantly enriched GO classes.

6.3.7 Expression of maternal and zygotic genes in zebrafish and *Tetraodon*

I wanted to compare the expression abundance of transcripts of maternal and zygotic origin between zebrafish and *Tetraodon*. The intent was to estimate the similarity or difference in the transcriptional repertoire of the teleost fishes during embryogenesis. I considered three lists representing maternal, maternal/zygotic

and zygotic genes in zebrafish and their expression in five early developmental stages as my starting dataset from a recently published study (Harvey et al., 2013). Only the coding genes were considered for this analysis (1677 maternal, 3055 maternal zygotic, 264 zygotic). Further, I obtained the predicted *Tetraodon* orthologs for the genes of my starting dataset. I could map 80% of zebrafish genes from the maternal list, 82% from the maternal/zygotic list and 63% from the zygotic list to their corresponding ortholog in *Tetraodon*. It is important to note that the *Tetraodon* 24 hpf stage is comparable to the beginning of somitogenesis in zebrafish (10.5 hpf). The zebrafish maternal genes show maximum expression at the 64 cell stage followed by gradual decrease in the high, shield and 90% epiboly stages (**Figure 6.10 A**). The *Tetraodon* orthologs of the zebrafish maternal list follow a similar course with the maximum average expression level in the eggs and a decrease in expression abundance at 30% epiboly compared to eggs. However, there is a slight increase in expression at 24 hpf compared to 30 epi, which suggests that some of the *Tetraodon* orthologs might also be expressed as zygotic transcripts post-degradation of the maternal genes. I found that 26% of the *Tetraodon* orthologs to zebrafish maternal genes show such behavior. This is a reflection of the fact that the zebrafish data were selected as being *maternally-expressed*, but not necessarily *maternal-specific* like the ones I selected in *Tetraodon*. Majority of the genes belonging to the maternal/zygotic list in zebrafish are highly expressed in all the sampled staged (**Figure 6.10 B**). Their corresponding *Tetraodon* orthologs also show a similar pattern. The zebrafish zygotic genes show a stark correlation in their expression abundance with their *Tetraodon* orthologs. In zebrafish from close

to negligible transcription the zygotic genes show an increased expression in high, shield and 90% epiboly, a pattern closely emulated by their *Tetraodon* counterparts (**Figure 6.10 C**). Thus the *Tetraodon* orthologs to the zebrafish maternal and zygotic genes show a similar expression variation during development.

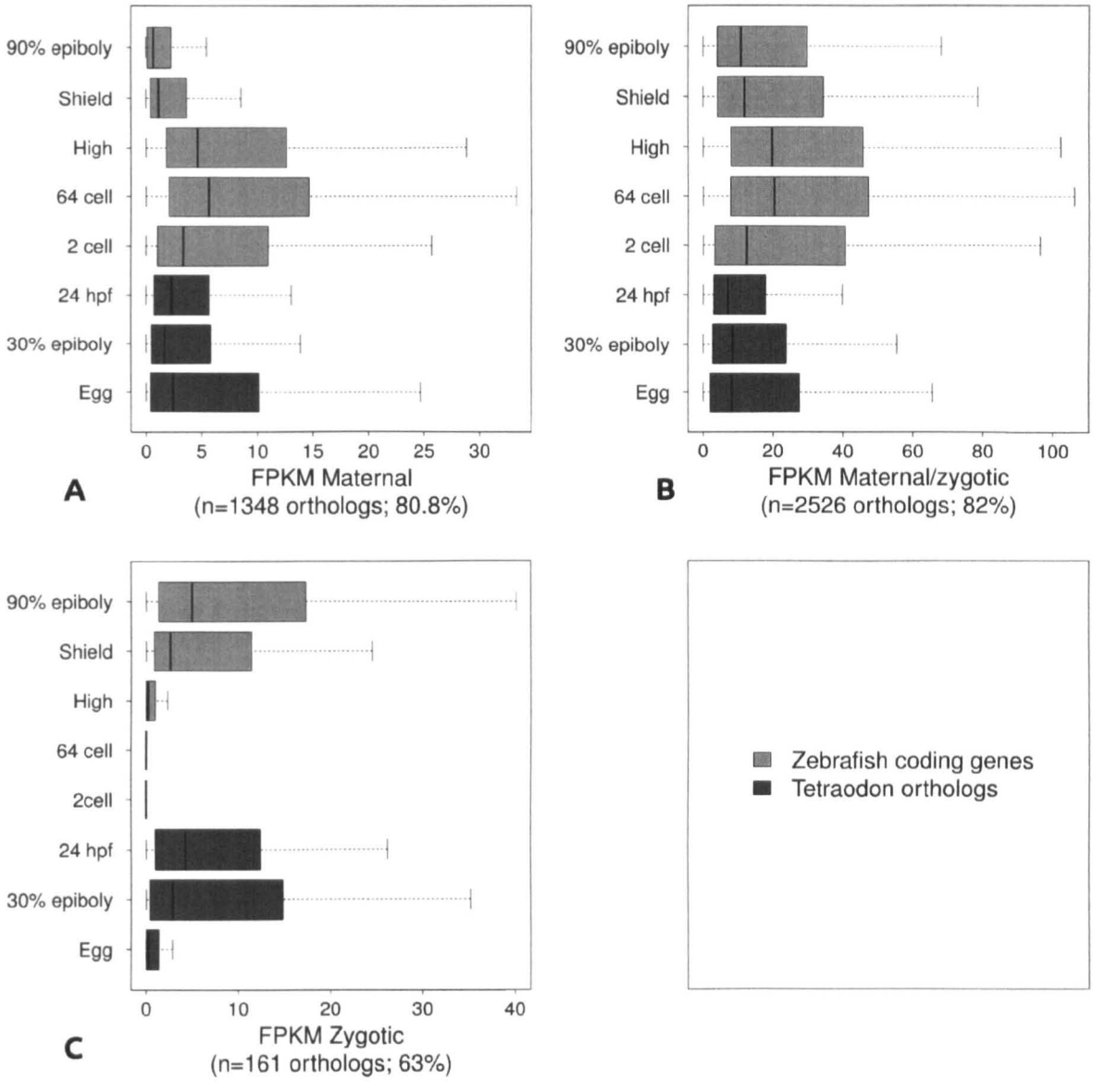


Figure 6.10 Expression abundance of zebrafish genes and their *Tetraodon* orthologs during development **A)** Maternal **B)** Maternal/Zygotic **C)** Zygotic

Further, I wanted to compare the expression of genes predicted to be maternal and

embryonic specific in *Tetraodon* with their orthologs in zebrafish. Since these genes are differentially expressed between the given developmental stages in *Tetraodon* they represent a select subset of genes which may be involved in distinguished cellular processes during MZT. I was able to map 85% of maternal and 82% of zygotic coding genes of *Tetraodon* to their zebrafish orthologs. The *Tetraodon* maternal specific genes are predominantly expressed in the egg with almost no detectable expression in the 30% epiboly and 24 hpf (**Figure 6.11 A**). The expression pattern of the zebrafish orthologs further support these genes to be maternal in origin, since most of them appear to be expressed in the 2 cell, 64 cell and high stages followed by a sudden decrease in transcriptional levels in the shield and 90% epiboly. Similar observation is made on the *Tetraodon* zygotic gene list where the zebrafish orthologs mimic the expression dynamics during development (**Figure 6.11 B**). The *Tetraodon* zygotic specific genes initiate expression at the 30% epiboly while their zebrafish coutnerparts are observed to be expressed minimally in the 2 cell and 64 cell stage followed by a steep rise of expression abundance in high, shield and 90% epiboly.

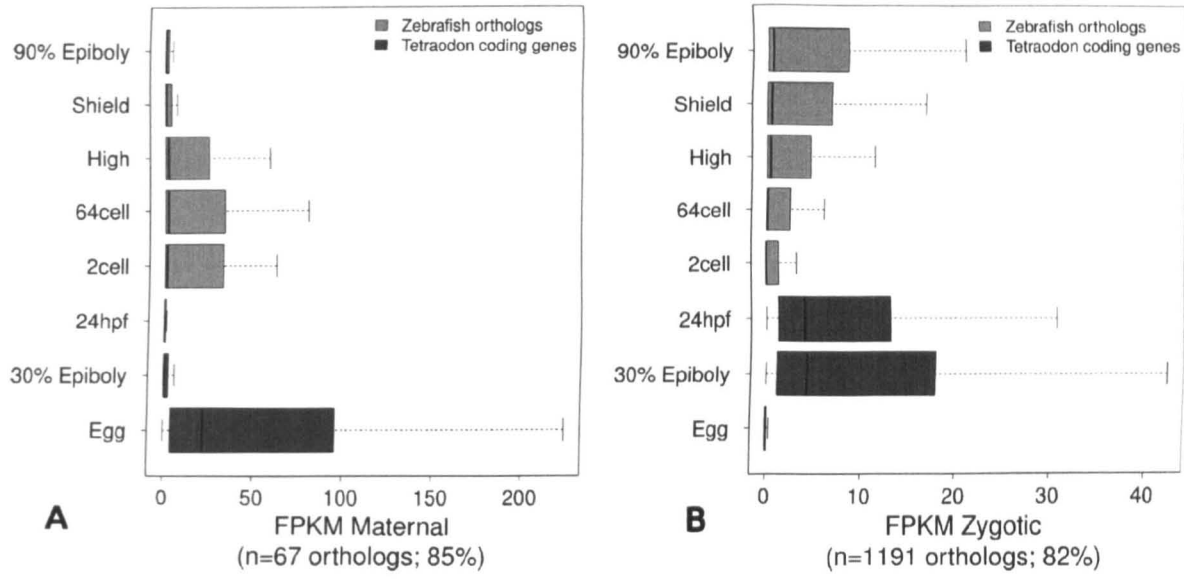


Figure 6.11 Expression abundance of *Tetraodon* genes and their zebrafish orthologs during development **A)** Maternal **B)** Zygotic

Thus the gene lists in *Tetraodon* derive support from the developmental expression of their zebrafish orthologs, to be representative of maternal and zygotic specific transcription. The conservation at the level of expression dynamics further indicates a potential functional retention of genes involved in various biological processes during MZT in teleost fishes.

6.3.8 Prediction of putative microsyntenic lincRNAs between *Tetraodon* and zebrafish

Finally, I decided to focus on identifying interesting lincRNA candidates which may be implicated in regulation of genes or pathways during *Tetraodon* embryogenesis. A major obstacle towards this aim is a lack of functional annotation for lincRNAs. The primary reason is their low level of sequence conservation which hinders the identification of lincRNA homologs across species

and hence their further association to a function by experimental validation. I have developed a pipeline (SynLinc) to compare the conserved microsyntenic lincRNAs (**Figure 4.1**). I have previously predicted a set of Vertebrate Microsyntenic LincRNAs (VMLs) with conserved microsynteny in human, mouse and zebrafish using the SynLinc pipeline. Here I used the pipeline to compare the lincRNAs predicted in the *Tetraodon* early developmental transcriptome with zebrafish lincRNAs predicted by two previously published studies and the Ensembl transcript annotation pipeline (Flicek et al., 2012b; Pauli et al., 2011a; Ulitsky et al., 2011). The pipeline predicted close to 800 lincRNA transcripts (zebrafish: 788 transcripts from 523 loci; *Tetraodon*: 796 transcripts from 667 loci) to show conserved microsynteny based upon the homology of a flanking coding gene. Further, I compared the predicted microsyntenic lincRNAs of *Tetraodon* with zebrafish VMLs to get 142 *Tetraodon* lincRNAs which have a predicted ortholog in human, mouse and zebrafish according to the SynLinc pipeline. I compared the proximal coding genes of all lincRNAs (10KB upstream and downstream) against the proximal coding genes of these 142 lincRNAs predicted to be microsyntenic in the four vertebrates (**Figure 6.12**). Several classes resulted significantly enriched among the 142 lincRNAs. Of these, I found very interesting the GO classes *histone lysine methylation, regulation of transcription* and *eye development* to be significantly enriched in the coding genes lying proximal to the microsyntenic lincRNAs. In the past lncRNAs have been reported to be involved in regulation of transcription and chromatin modifications (Batista and Chang, 2013). Further long non-coding RNAs are known to be expressed near coding genes which are implicated in regulation of

cellular development and differentiation specially in the brain (Aprea et al., 2013; Guttman et al., 2011; Laurent et al., 2013). However a recent report has mentioned the presence of a select set of lincRNAs which are specifically expressed in the human eye and are conserved in sequence and expression pattern in other mammals, thus these candidate lincRNAs are expected to play an important role in mammalian eye development (Mustafi et al., 2013). Thus the putative microsyntenic lincRNAs are observed to show an enrichment to lie near coding genes involved in regulation of transcription, chromatin modification and eye development.

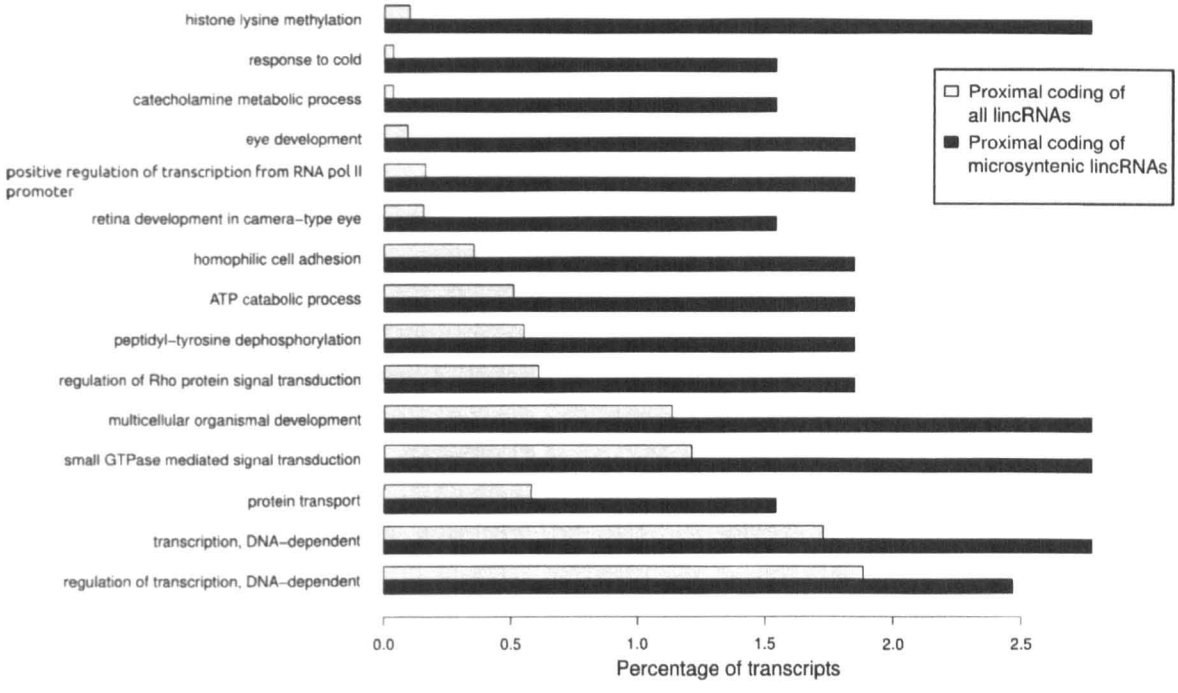


Figure 6.12 Gene ontology enrichment analysis for coding genes lying proximal to microsyntenic lincRNAs in *Tetraodon*. The x-axis represents the percentage of transcripts which are defined by a particular GO biological process and the y-axis represents the significantly enriched GO classes.

6.3.9 Comparison of the assembled transcript models with transcript models from Ensembl

The predicted coding transcripts represent 86% of known Ensembl coding gene models but 74% of the coding transcripts which map to an Ensembl gene ID show the presence of at least one novel exon or an alternative splice site thus being classified as a novel isoform of a known gene. Indeed alternative splicing events are partly responsible for the vertebrate transcriptional complexity (Barbosa-Morais et al., 2012; Braunschweig et al., 2013) but in specific cases the presence of additional exons may suggest a partial gene model being present in the reference data. A good example is that of the zebrafish Smaug 1 *(Smg1)* gene (12KB, 63 exons) whose *Tetraodon* ortholog is represented by a comparatively smaller gene (1KB, 7 exons). Manual inspection of *Tetradodon Smg1* genomic region led me to identify two transcript models (*TCONS_00047347*: 3KB, 22 exons, *TCONS_00047348*: 6KB, 35 exons) (**Figure 6.13 A**) which are co-linear to each other and are placed in the same transcribed locus by Cufflinks. The 3' end of *TCONS_00047347* lies at a distance of 50 bases from the 5' end of the *TCONS_00047348*. To check whether the two transcripts are not isoforms but part of the same transcript representing the *Smg1* gene I concatenated their sequences and did a tBLASTx search against the zebrafish cDNA sequences in the Ensembl database. Indeed the concatenated sequence shows homology with 27 exons of zebrafish *Smg1* gene (**Figure 6.13 B**). The evidence from the assembled transcripts, and their homology search against the zebrafish cDNA sequences indicates that the Ensembl *Tetraodon Smg1* gene is represented by an inadequate gene model.

Such an example highlights the additional utility of the assembled gene models to improve upon the existing *Tetraodon* gene annotations.

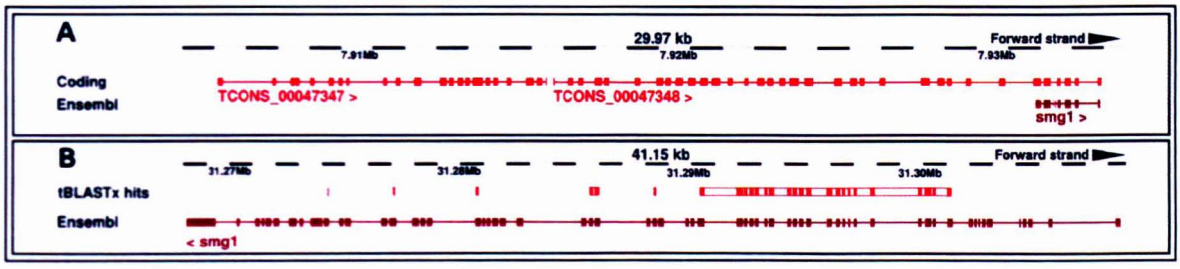


Figure 6.13 Ensembl genome browser screenshot of *Smg1* gene in **A)** *Tetraodon* **B)** zebrafish. The Ensembl track represents the gene models from Ensembl database. The Coding track the assembled coding gene models. The tBLASTx hits track shows the aligned regions of the concatenated *Tetraodon* transcripts (*TCONS_00047347, TONS_00047348*) against the zebrafish *Smg1* cDNA sequence obtained by tBLASTx comparison.

6.4 Conclusion

In the current study I have tried to understand the molecular basis of the developmental processes during embryogeneis in *Tetraodon*. The transcriptome data produced in this study shows a high coverage and depth harnessing the power of high-throughput sequencing technologies. I performed extensive bioinformatic analyses to assemble and annotate the developmental transcriptome and assess the temporal variation in transcript abundance in relation to their functional associations. High throughput sequencing of polyadenylated RNA across three developmental stages of *Tetraodon* (fertilised egg, 30% epiboly and 24 hours post fertilisation) resulted in the assembly of 53,543 coding transcripts from 19,414 loci (representing 86% of annotated Ensembl *Tetraodon* genes) and 4026 long

non-coding transcripts from 3508 loci. I found about 16% (16% of coding and 16% of long non-coding) of all transcripts to show a decrease in transcriptional abundance from the egg to 24 hours post fertilisation (24 hpf), thus suggesting them to be of maternal origin, a few of them known to be important regulators of fertilisation and cell division. The added advantage comes from the identification of the lncRNAs expressed during development which show the cellular non-coding diaspora to be actively involved in the process of embryogenesis. The differentially expressed maternal and embryonic transcripts are shown to be representative of processes and pathways which are in agreement with their known functional roles during embryonic development. Further, genes reported to be of maternal and zygotic origin in zebrafish show a similar expression dynamics in *Tetraodon* and *vice-versa*. Thus, even though there is a lack of replicates for the RNAseq experiment, evidences from gene ontology analyses and comparison of expression abundances support the integrity of the mapping and assembly of the early developmental transcriptome in *Tetraodon*. I found about 800 *Tetraodon* lincRNAs with conserved position compared to published zebrafish lincRNAs and 142 of them are predicted to be syntenic with other vertebrates, suggesting these candidates to be transcriptionally linked to their genomic neighborhood. Finally, the current study provides a well annotated assembly of early developmental transcripts in *Tetraodon*, a significant percentage representing novel isoforms of known genes. My analysis resulted in the identification of ~3000 protein-coding loci previously unreported in *Tetraodon*. Hence the assembled transcripts are expected to aid in improving the existing gene models as well as documenting

novel isoforms of known genes.

Chapter 7

General conclusion and future directions

7.1 The conservation factor in long non-coding RNAs

My Ph.D. project has been focused towards gaining insights into the evolution, the structure and the functions of lncRNAs, through computational approaches and the usage of large scale functional genomics data. Historically, long non-coding RNAs (lncRNAs) have not been associated with conservation of function or feature across a diverse range of species and this made my project challenging. However, there have been a few reports which mention presence of conserved chromatin signature (Guttman et al., 2009), transcription pattern (Kutter et al., 2012; Managadze et al., 2013) and sequence (Mustafi et al., 2013; Ponjavic et al., 2009; Young et al., 2012) in lncRNAs. However, majority of such studies have not considered species separated by long evolutionary distances, mostly remaining within the realm of mammalian genomes. Nevertheless lncRNAs are repeatedly shown to be involved in regulation of important biological processes, specially development of brain (Aprea et al., 2013; Lipovich et al., 2013; Mercer et al., 2008) and differentiation of pluripotent cells (Dinger et al., 2008; Guttman et al., 2011; Lin et al., 2011b; Pauli et al., 2011a). A question raised is, whether the lack of evolutionary conservation may be considered as an absence of function in the

majority of lncRNAs. The answer to this question lies within the following issues, which currently do not have a clear understanding

- How do we define a transcript as a lncRNA?
- What is the correct estimate of the transcribed lncRNA population in an organism at a given developmental stage or tissue?
- What are the properties associated with lncRNAs, which share similar functions in different organisms and how can we isolate them?

Although extensive experimental validation is necessary to arrive at a definite conclusion, it is also important to obtain a computational perspective to the issues mentioned above. During the tenure of my Ph.D. I have made an attempt to answer these questions by developing computational protocols for identification of lncRNAs and estimation of their conservation. I have further used these protocols to analyse large scale sequencing datasets from a specific tissue and from early developmental stages, to predict lncRNAs which may play an important functional role in a metabolic disorder or during embryogenesis.

7.2 Computational prediction of lncRNAs

There are numerous reports of large scale prediction of lncRNAs from next-generation sequencing data (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2011a) but a common computational framework for lncRNA is lacking. Homology to coding genes, lack of a long open reading frame and a low protein coding potential (based on coding frequency and nucleotide composition) are the

principal measures used to predict lncRNAs. But none of the previously published reports have defined a standard protocol which can be universally applied and compared to predict lncRNAs in different organisms. Hence I developed a computational pipeline which can annotate both coding and long non-coding sequences in a large dataset. I developed the pipeline (Annocript) principally to predict lncRNAs on different RNAseq datasets I analysed during the period of my Ph.D. However, I also plan to make the pipeline available to the scientific community as a resource for annotating coding and non-coding transcriptomes in diverse organisms. Apart from predicting lncRNAs the principal advantage of Annocript is its ability to make optimal use of parallel processing to achieve a significant loss in annotation time. The pipeline is able to accurately identify known protein-coding sequences. I have tested the pipeline on previously published lncRNA datasets (human and zebrafish) and the results suggests that a sizable fraction of the reported lncRNAs might be actually coding for proteins or small peptides. This fact highlights the issue of defining the optimal lncRNAome size in an organism. I believe that our current knowledge is limited to accurately predict the correct number of lncRNAs. However, a standard measure for lncRNA prediction may help avoid over-estimation of lncRNA population in a given RNA sample. Further it will aid in making comparisons between different studies which will employ a similar tool to predict lncRNAs. Hence a pipeline like Annocript will prove to be highly useful in all future studies which aim towards identification of lncRNAs in a given cell, tissue or organism.

7.3 Sequence conservation in lncRNAs: Short segments in a small population

Additional measures of conservation could be used to associate a putative functionality to predicted lncRNA candidates. Hence, I defined a computational pipeline which can be effectively employed on diverse organisms to measure the sequence conservation of lncRNAs (Basu et al., 2013). Specifically a combination of two parameters (BLAST e-value and alignment length) were deemed sufficient to select candidate lncRNAs which show the presence of conserved sequence motifs between species separated by a large evolutionary distance (mouse and zebrafish). However, a very small percentage of the my initial dataset could be predicted as conserved. Further, if I consider only those regions of conservation which lie exclusively in intergenic or intronic regions (with respect to protein coding genes) then I have a set of ~50 mouse lncRNAs which show a significant level of sequence conservation with zebrafish. It is important to note that this small dataset is potentially interesting to experimentally validate and may provide novel insights into the mechanism of functionally conserved lncRNAs. The general lack of sequence conservation in lncRNAs led me consider an alternative approach to associate a lincRNA with a putative function.

7.4 Microsynteny in lncRNAs

The retention of position in lncRNAs is an approach which does not presume an existing conservation of sequence or secondary structure to associate putative

homology. I developed a pipeline to measure the positional conservation of long intergenic non-coding RNAs (lincRNAs) between different species (SynLinc). The pipeline was used to predict a set of lincRNAs which are observed to retain their position with respect to a flanking coding gene in human, mouse and zebrafish. I address these lincRNAs as Vertebrate Microsyntenic LincRNAs (VMLs). I expected the retention of position to be a function of a shared regulation or co-regulation between the lincRNAs and its nearby protein coding gene. But I failed to observe significant co-expression patterns between the majority of lincRNAs and their flanking coding genes in human and mouse. However, I found a subset (15%) of the conserved lincRNAs to show a significant expression correlation with their flanking genes during early embryonic development in zebrafish. Interestingly, the sequence space between a VML and its flanking orthologous coding genes shows an enrichment for sequence conservation thus arguing against the presence of the lincRNA due to a random evolutionary event. Further, I have observed an enrichment of the human and zebrafish VMLs to lie near an active conserved enhancer, while this enrichment is not observed in mouse. A recent finding has demonstrated that intergenic lincRNAs in mouse can be divided into classes based on the overlap of chromatin features (Marques et al., 2013). Further, the study states that lincRNAs which are associated with enhancers do not differ structurally from other lincRNAs, but are less conserved (sequence) and tend to be co-expressed with their proximal coding genes. The results from this finding illustrates the fact, that lincRNAs can be considered as dynamic molecules further divided into multiple classes with diverse functions. I have predicted a set of

lincRNAs which retain their position across evolution but I did not find any evidence implicating such lincRNAs to be functionally associated with their flanking coding genes. Yet, this is a work in progress and I started to use the lncRNA information from another teleost fish to further refine my results and identify functionally conserved lincRNAs.

7.5 Prediction of islet cell specific lincRNAs in zebrafish

I generated the zebrafish islet cell lincRNA catalog and predicted a potential list of lincRNA candidates which may be involved in pancreatic development and differentiation of islet cells. The hypothesis is such lincRNAs may be involved in regulation of genes or pathways implicated in type 2 diabetes. I have been able to isolate two major issues during identification of lincRNAs, related to the analysis of the RNAseq data. The first issue is multiple mapping of short reads on the genome and the second issue is the assembly of reads into transcripts. I have demonstrated that slight parametric changes in mapping short reads on the genome are magnified in the downstream assembly of transcripts, often leading to spurious transcript models and differences in the number of transcripts assembled. Further, I also observed that such differences are more prominent for intergenic or long non-coding RNA transcripts hence again raising the previous issue of, how to reliably estimate a lncRNA population in a RNA sample computationally. Hence, I defined a mapping and assembly protocol which takes advantage of the inherent features of the Tophat (Kim et al., 2013a) and Cufflinks (Trapnell et al., 2012)

softwares to reduce transcript mis-assembly. Further I used the Annocript and SynLinc pipelines to define a set of lincRNAs significantly upregulated in the zebrafish islet cells. Currently, experimental validations are being carried out for these candidate lincRNAs in the laboratory of my external supervisor Dr Ferenc Müller in the University of Birmingham, UK.

7.6 LncRNAs in embryogenesis

Finally, I have analysed the early developmental transcriptome of the spotted green pufferfish (*Tetraodon nigroviridis*), to predict coding and long non-coding transcripts transcribedd during early embryogenesis. I mapped the expression of the coding and the long non-coding transcripts globally (**Figure 6.4**) to observe a dynamic expression of the lncRNAs in different genomic loci, sometimes concurrent with coding genes. Further, based on GO enrichment analysis and expression comparison with zebrafish orthologs, I demonstrated the similarity between the predicted maternal and embryonic specific coding transcripts in *Tetraodon* with the defined roles of such transcripts in other vertebrates. It is interesting to note that the early developmental lincRNAs in *Tetraodon* are observed to lie near coding genes which are implicated in development of the brain and chromatin modifications, two functions with which lincRNAs are widely associated. Further, I used the SynLinc pipeline to compare the *Tetraodon* lincRNAs with zebrafish lincRNAs as well as VMLs. The *Tetraodon* lincRNAs which show positional conservation with VMLs are enriched to lie near coding genes

implicated in processes like eye development and regulation of transcription. Previously, another study has demonstrated a dynamic population of lncRNAs being transcribed in zebrafish during embryogenesis (Pauli et al., 2011a). Here, I report a large set of lncRNAs in another teleost fish also showing a robust expression pattern during early development. Currently this is a work in progress and I will further make an in-depth comparison of the lincRNA population of zebrafish and *Tetraodon* to gain further insight into their possible duplication, enrichment of chromatin marks and sequence conservation.

7.7 Future perspectives

A major period of my PhD tenure was spent on defining sensitive protocols for the prediction of lncRNAs and their comparative analysis between different species. I have predicted lncRNAs from two different RNAseq datasets, highlighting the various computational hurdles which come across during such a task. Currently I have a set of well defined computational protocols to predict and compare lncRNAs, whose merit has been demonstrated in different analyses. I aim to make these pipelines available to the scientific community. Further, I want to continue with the analysis of the *Tetraodon* early developmental lncRNAome. Specifically I plan to compare intra and inter species lincRNA microsynteny between zebrafish and *Tetraodon* and associate the results with various parameters like sequence conservation, presence of transposable elements, chromatin state and developmental expression pattern. Considering the fact that lincRNAs are

associated independently with the given parameters in various published reports, I believe that such an analysis will provide the basic framework to assign a function with respect to lincRNA conservation. Further comparison of the results with mammalian lincRNAs may provide valuable insight into the evolution lincRNAs and their function.

References

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G.P., Lim, A.Y.M., Hajan, H.S., Collas, P., Bourque, G., et al. (2011). Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. *21*, 1328–1338.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J.C., Suzuki, H., Daub, C.O., Hayashizaki, Y., and Lenhard, B. (2009). Transcriptional features of genomic regulatory blocks. Genome Biol. *10*, R38.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Amaral, P.P., Dinger, M.E., Mercer, T.R., and Mattick, J.S. (2008). The eukaryotic genome as an RNA machine. Science *319*, 1787–1789.

Amaral, P.P., Neyt, C., Wilkins, S.J., Askarian-Amiri, M.E., Sunkin, S.M., Perkins, A.C., and Mattick, J.S. (2009). Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. RNA N. Y. N *15*, 2013–2027.

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., and Mattick, J.S. (2011). lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. *39*, D146–151.

Aoyama, M., Ozaki, T., Inuzuka, H., Tomotsune, D., Hirato, J., Okamoto, Y., Tokita, H., Ohira, M., and Nakagawara, A. (2005). LMO3 interacts with neuronal transcription factor, HEN2, and acts as an oncogene in neuroblastoma. Cancer Res. *65*, 4587–4597.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science *297*, 1301–1310.

Aprea, J., Prenninger, S., Dori, M., Ghosh, T., Monasor, L.S., Wessendorf, E., Zocher, S., Massalini, S., Alexopoulou, D., Lesche, M., et al. (2013). Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment. EMBO J. *advance online publication*.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*, D115–119.

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. (2002). Gene expression during the life cycle of Drosophila melanogaster. Science *297*, 2270–2275.

Van Arensbergen, J., García-Hurtado, J., Moran, I., Maestro, M.A., Xu, X., Van de Casteele, M., Skoudy, A.L., Palassini, M., Heimberg, H., and Ferrer, J. (2010). Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. Genome Res. *20*, 722–732.

Arrial, R., Togawa, R., and Brigido, M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. BMC Bioinformatics *10*, 239.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. *40*, W597–603.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Aston-Mourney, K., Zraika, S., Udayasankar, J., Subramanian, S.L., Green, P.S., Kahn, S.E., and Hull, R.L. (2013). Matrix metalloproteinase-9 reduces islet amyloid formation by degrading islet amyloid polypeptide. J. Biol. Chem. *288*, 3553–3559.

Azumi, K., Sabau, S.V., Fujie, M., Usami, T., Koyanagi, R., Kawashima, T., Fujiwara, S., Ogasawara, M., Satake, M., Nonaka, M., et al. (2007). Gene expression profile during the life cycle of the urochordate Ciona intestinalis. Dev. Biol. *308*, 572–582.

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. (2001). Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics *78*, 73–82.

Van Bakel, H., and Hughes, T.R. (2009). Establishing legitimacy and function in the new transcriptome. Brief. Funct. Genomic. Proteomic. *8*, 424–436.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol. *8*, e1000371.

Balwierz, P.J., Carninci, P., Daub, C.O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., and van Nimwegen, E. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. *10*, R79.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr, Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. *22*, 1646–1657.

Bao, L., Zhou, M., and Cui, Y. (2008). CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. Nucleic Acids Res. *36*, D83–D87.

Barber, B.A., and Rastegar, M. (2010). Epigenetic control of Hox genes during neurogenesis, development, and disease. Ann. Anat. Anat. Anz. Off. Organ Anat. Ges. *192*, 261–274.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science *338*, 1587–1593.

Barry, G., and Mattick, J.S. (2012). The role of regulatory RNA in cognitive evolution. Trends Cogn. Sci. *16*, 497–503.

Basu, S., Müller, F., and Sanges, R. (2013). Examples of sequence conservation analyses capture a subset of mouse long non-coding RNAs sharing homology with fish conserved genomic elements. BMC Bioinformatics *14*, S14.

Batista, P.J., and Chang, H.Y. (2013). Long Noncoding RNAs: Cellular Address Codes in Development and Disease. Cell *152*, 1298–1307.

Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Dev. Camb. Engl. *130*, 889–900.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N.P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., et al. (2012). Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell *24*, 3949–3965.

Becker, T.S., and Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: a matter of resolution. Mol. Genet. Genomics *278*, 487–491.

Becker, T.S., and Rinkwitz, S. (2012). Zebrafish as a genomics model for human neurological and polygenic disorders. Dev. Neurobiol. *72*, 415–428.

Bedell, V.M., Westcot, S.E., and Ekker, S.C. (2011). Lessons from morpholino-based screening in zebrafish. Brief. Funct. Genomics *10*, 181–188.

Begum, F., Ghosh, D., Tseng, G.C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res. *40*, 3777–3784.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E., and Goodman, H.M. (1980). Sequence of the human insulin gene. Nature *284*, 26–32.

Berghoff, E.G., Clark, M.F., Chen, S., Cajigas, I., Leib, D.E., and Kohtz, J.D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development dev.099390.

Bernard, D., Prasanth, K.V., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M.Q., Sedel, F., Jourdren, L., Coulpier, F., et al. (2010). A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J. *29*, 3082–3093.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Blackburn, P.R., Campbell, J.M., Clark, K.J., and Ekker, S.C. (2013). The CRISPR system--keeping zebrafish gene targeting fresh. Zebrafish *10*, 116–118.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Res. *14*, 708–715.

Blelloch, R., and Gutkind, J.S. (2013). Epigenetics, noncoding RNAs, and cell

signaling — crossroads in the regulation of cell fate decisions. Curr. Opin. Cell Biol. *25*, 149–151.

Bogdanovic, O., Fernandez-Miñán, A., Tena, J.J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S.J., Veenstra, G.J.C., and Gómez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. *22*, 2043–2053.

Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The product of the H19 gene may function as an RNA. Mol. Cell. Biol. *10*, 28–36.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. (2003). Sequence information can be obtained from single DNA molecules. Proc. Natl. Acad. Sci. U. S. A. *100*, 3960–3964.

Braun, M., Ramracheya, R., and Rorsman, P. (2012). Autocrine regulation of insulin secretion. Diabetes Obes. Metab. *14 Suppl 3*, 143–151.

Braunschweig, U., Gueroussov, S., Plocik, A.M., Graveley, B.R., and Blencowe, B.J. (2013). Dynamic Integration of Splicing within Gene Regulatory Pathways. Cell *152*, 1252–1269.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature *366*, 265–268.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Brugman, S., Liu, K.-Y., Lindenbergh-Kortleve, D., Samsom, J.N., Furuta, G.T., Renshaw, S.A., Willemsen, R., and Nieuwenhuis, E.E.S. (2009). Oxazolone-induced enterocolitis in zebrafish depends on the composition of the intestinal microbiota. Gastroenterology *137*, 1757–1767.e1.

Burgess, D.J. (2011). Non-coding RNA: HOTTIP goes the distance. Nat. Rev. Genet. *12*, 300.

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S., and Bilofsky, H.S. (1985). The GenBank nucleic acid sequence database. Comput. Appl. Biosci. CABIOS *1*, 225–233.

Burnicka-Turek, O., Mohamed, B.A., Shirneshan, K., Thanasupawat, T., Hombach-Klonisch, S., Klonisch, T., and Adham, I.M. (2012). INSL5-deficient mice display an alteration in glucose homeostasis and an impaired fertility. Endocrinology *153*, 4655–4665.

Burns, C.G., and MacRae, C.A. (2006). Purification of hearts from zebrafish embryos. BioTechniques *40*, 274, 276, 278 passim.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915 –1927.

Calderon, M.R., Verway, M., An, B.-S., DiFeo, A., Bismar, T.A., Ann, D.K., Martignetti, J.A., Shalom-Barak, T., and White, J.H. (2012). Ligand-dependent corepressor (LCoR) recruitment by Kruppel-like factor 6 (KLF6) regulates expression of the cyclin-dependent kinase inhibitor CDKN1A gene. J. Biol. Chem. *287*, 8662–8674.

Calin, G.A., Liu, C., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E.J., Wojcik, S.E., et al. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell *12*, 215–229.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Camargo, A.A., Samaia, H.P., Dias-Neto, E., Simão, D.F., Migotto, I.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A., et al. (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc. Natl. Acad. Sci. U. S. A. *98*, 12103–12108.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat. Rev. Genet. *14*, 100–112.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499–509.

Ceribelli, M., Alcalay, M., Viganò, M.A., and Mantovani, R. (2006). Repression of new p53 targets revealed by ChIP on chip experiments. Cell Cycle Georget. Tex *5*, 1102–1110.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Cha, S.-W., Tadjuidje, E., White, J., Wells, J., Mayhew, C., Wylie, C., and Heasman, J. (2009). Wnt11/5a complex formation caused by tyrosine sulfation increases canonical signaling activity. Curr. Biol. CB *19*, 1573–1580.

Chang, L.-Y., Harduin-Lepers, A., Kitajima, K., Sato, C., Huang, C.-J., Khoo, K.-H., and Guérardel, Y. (2009). Developmental regulation of oligosialylation in zebrafish. Glycoconj. J. *26*, 247–261.

Cheetham, S.W., Gruhl, F., Mattick, J.S., and Dinger, M.E. (2013). Long noncoding RNAs and the genetics of cancer. Br. J. Cancer *108*, 2419–2425.

Chen, L.-L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding

RNA. Mol. Cell *35*, 467–478.

Chen, X., and Yan, G.-Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics *29*, 2617–2624.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013a). LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. *41*, D983–986.

Chen, Y.-C., Fueger, P.T., and Wang, Z. (2013b). Depletion of PAK1 enhances ubiquitin-mediated survivin degradation in pancreatic β-cells. Islets *5*, 22–28.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. *22*, 1658–1667.

Chew, G.-L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development *140*, 2828–2834.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnár, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. *11*, R72.

Chooniedass-Kothari, S., Emberley, E., Hamedani, M.K., Troup, S., Wang, X., Czosnek, A., Hube, F., Mutawe, M., Watson, P.H., and Leygue, E. (2004). The steroid receptor RNA activator is the first functional RNA encoding a protein. FEBS Lett. *566*, 43–47.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., et al. (2011). The Reality of Pervasive Transcription. PLoS Biol *9*, e1000625.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res. *22*, 885–898.

Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J. Cell Biol. *132*, 259–275.

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell *33*, 717–726.

Cloonan, N., and Grimmond, S.M. (2008). Transcriptome content and dynamics at single-nucleotide resolution. Genome Biol. *9*, 234.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*, 613–619.

Collins, J.E., White, S., Searle, S.M.J., and Stemple, D.L. (2012). Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res. *22*, 2067–2078.

Collombat, P., Xu, X., Ravassard, P., Sosa-Pineda, B., Dussaud, S., Billestrup, N., Madsen, O.D., Serup, P., Heimberg, H., and Mansouri, A. (2009). The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into alpha and subsequently beta

cells. Cell *138*, 449–462.

Conley, A.B., and Jordan, I.K. (2012). Epigenetic regulation of human cis-natural antisense transcripts. Nucleic Acids Res. *40*, 1438–1445.

Cornide-Petronio, M.E., and Barreiro-Iglesias, A. (2013). Role of Slit and Robo proteins in the development of dopaminergic neurons. Dev. Neurosci. *35*, 285–292.

Costa, V., Aprile, M., Esposito, R., and Ciccodicola, A. (2013). RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur. J. Hum. Genet. *21*, 134–142.

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., and Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics *14*, 648.

Cruveiller, S., Clay, O., Jabbari, K., and Bernardi, G. (2007). Simple proteomic checks for detecting noncoding RNA. Proteomics *7*, 361–363.

Dash, S.K. (2013). Cognitive impairment and diabetes. Recent Pat. Endocr. Metab. Immune Drug Discov. *7*, 155–165.

Dearry, A., Gingrich, J.A., Falardeau, P., Fremeau, R.T., Jr, Bates, M.D., and Caron, M.G. (1990). Molecular cloning and expression of the gene for a human D1 dopamine receptor. Nature *347*, 72–76.

DeMare, L.E., Leng, J., Cotney, J., Reilly, S.K., Yin, J., Sarro, R., and Noonan, J.P. (2013). The genomic landscape of cohesin-associated chromatin interactions. Genome Res. *23*, 1224–1234.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*, 578–582.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Devaskar, S.U., Giddings, S.J., Rajakumar, P.A., Carnaghi, L.R., Menon, R.K., and Zahm, D.S. (1994). Insulin gene expression and insulin synthesis in mammalian neuronal cells. J. Biol. Chem. *269*, 8445–8454.

Dimitrieva, S., and Bucher, P. (2012). UCNEbase–a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res. *41*, D101–D109.

Ding, G.-L., Wang, F.-F., Shu, J., Tian, S., Jiang, Y., Zhang, D., Wang, N., Luo, Q., Zhang, Y., Jin, F., et al. (2012). Transgenerational glucose intolerance with Igf2/H19 epigenetic alterations in mouse islet induced by intrauterine hyperglycemia. Diabetes *61*, 1133–1142.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Soldà, G., Simons, C., et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. *18*, 1433–1445.

Dinger, M.E., Gascoigne, D.K., and Mattick, J.S. (2011). The evolution of RNAs with multiple functions. Biochimie *93*, 2013–2018.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. Proc. Natl. Acad. Sci. U. S. A. *110*, 5294–5300.

Doyon, Y., McCammon, J.M., Miller, J.C., Faraji, F., Ngo, C., Katibah, G.E., Amora, R., Hocking, T.D., Zhang, L., Rebar, E.J., et al. (2008). Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. Nat. Biotechnol. *26*, 702–708.

Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R., et al. (2012). The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res. *40*, D918–923.

Driever, W., Solnica-Krezel, L., Schier, A.F., Neuhauss, S.C., Malicki, J., Stemple, D.L., Stainier, D.Y., Zwartkruis, F., Abdelilah, S., Rangini, Z., et al. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. Development *123*, 37–46.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006a). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science *312*, 1653–1655.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006b). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science *312*, 1653–1655.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma. Oxf. Engl. *21*, 3439–3440.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat. Cell Biol. *10*, 1106–1113.

Ebralidze, A.K., Guibal, F.C., Steidl, U., Zhang, P., Lee, S., Bartholdy, B., Jorda, M.A., Petkova, V., Rosenbauer, F., Huang, G., et al. (2008). PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. Genes Dev. *22*, 2085–2092.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., et al. (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA Biol. *9*, 1076–1087.

Ellis, B.C., Molloy, P.L., and Graham, L.D. (2012). CRNDE: a long non-coding RNA involved in CanceR, Neurobiology, and DEvelopment. Front. Non-Coding RNA *3*, 270.

Engström, P.G., Ho Sui, S.J., Drivenes, Ø., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. *17*, 1898–1908.

Eun, B., Sampley, M.L., Good, A.L., Gebert, C.M., and Pfeifer, K. (2013). Promoter cross-talk via a shared enhancer explains paternally biased expression of Nctc1 at the Igf2/H19/Nctc1 imprinted locus. Nucleic Acids Res. *41*, 817–826.

Faghihi, M.A., and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. Nat. Rev. Mol. Cell Biol. *10*, 637–643.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., 3rd, Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat. Med. *14*, 723–730.

Faghihi, M.A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M.P., Nalls, M.A., Cookson, M.R., St-Laurent, G., 3rd, and Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol. *11*, R56.

Falkmer, S. (1993). Phylogeny and ontogeny of the neuroendocrine cells of the gastrointestinal tract. Endocrinol. Metab. Clin. North Am. *22*, 731–752.

FANTOM Consortium, Suzuki, H., Forrest, A.R.R., van Nimwegen, E., Daub, C.O., Balwierz, P.J., Irvine, K.M., Lassmann, T., Ravasi, T., Hasegawa, Y., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat. Genet. *41*, 553–562.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. *41*, 563–571.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Mach. Learn. *31*, 1–38.

Fernandes, I., Bastien, Y., Wai, T., Nygard, K., Lin, R., Cormier, O., Lee, H.S., Eng, F., Bertos, N.R., Pelletier, N., et al. (2003). Ligand-dependent nuclear receptor corepressor LCoR functions by histone deacetylase-dependent and -independent mechanisms. Mol. Cell *11*, 139–150.

Flicek, P., Aken, B.L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., et al. (2010). Ensembl's 10th year. Nucleic Acids Res. *38*, D557–562.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. Nucleic Acids Res. *39*, D800–806.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012a). Ensembl 2012. Nucleic Acids Res. *40*, D84–90.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2012b). Ensembl 2013. Nucleic Acids Res. *41*, D48–55.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. Nucleic Acids Res. gkt1196.

Fraisl, P. (2013). Crosstalk between oxygen- and nitric oxide-dependent signaling pathways in angiogenesis. Exp. Cell Res. *319*, 1331–1339.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Gabory, A., Jammes, H., and Dandolo, L. (2010). The H19 locus: role of an imprinted non-coding RNA in growth and development. BioEssays News Rev. Mol. Cell. Dev. Biol. *32*, 473–480.

Gao, N., Le Lay, J., Qin, W., Doliba, N., Schug, J., Fox, A.J., Smirnova, O., Matschinsky, F.M., and Kaestner, K.H. (2010). Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. Mol. Endocrinol. Baltim. Md *24*, 1594–1604.

Gardiner, D.M., Blumberg, B., Komine, Y., and Bryant, S.V. (1995). Regulation of HoxA expression in developing and regenerating axolotl limbs. Dev. Camb. Engl. *121*, 1731–1741.

Gelling, R.W., Morton, G.J., Morrison, C.D., Niswender, K.D., Myers, M.G., Jr, Rhodes, C.J., and Schwartz, M.W. (2006). Insulin action in the brain contributes to glucose lowering during insulin treatment of diabetes. Cell Metab. *3*, 67–73.

Geng, Y.J., Xie, S.L., Li, Q., Ma, J., and Wang, G.Y. (2011). Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. J. Int. Med. Res. *39*, 2119–2128.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Géraudie, J., and Borday Birraux, V. (2003). Posterior hoxa genes expression during zebrafish bony fin ray development and regeneration suggests their involvement in scleroblast differentiation. Dev. Genes Evol. *213*, 182–186.

Gesell, T., and Washietl, S. (2008). Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics *9*, 248.

Giraldez, A.J. (2010). microRNAs, the cell's Nepenthe: clearing the past during the maternal-to-zygotic transition and cellular reprogramming. Curr. Opin. Genet. Dev. *20*, 369–375.

Goldsmith, J.R., and Jobin, C. (2012). Think Small: Zebrafish as a Model System of Human Pathology. J. Biomed. Biotechnol. *2012*, 817341.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature *470*, 284–288.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. *36*, 3420–3435.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2010). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

Grünert, S., and St Johnston, D. (1996). RNA localization and the development of asymmetry during Drosophila oogenesis. Curr. Opin. Genet. Dev. *6*, 395–402.

Guay, C., Jacovetti, C., Nesca, V., Motterle, A., Tugay, K., and Regazzi, R. (2012). Emerging roles of non-coding RNAs in pancreatic β-cell function and dysfunction. Diabetes Obes. Metab. *14*, 12–21.

Guérardel, Y., Chang, L.-Y., Maes, E., Huang, C.-J., and Khoo, K.-H. (2006). Glycomic survey mapping of zebrafish identifies unique sialylation pattern. Glycobiology *16*, 244–257.

Guil, S., Soler, M., Portela, A., Carrère, J., Fonalleras, E., Gómez, A., Villanueva, A., and Esteller, M. (2012). Intronic RNAs mediate EZH2 regulation of epigenetic targets. Nat. Struct. Mol. Biol. *19*, 664–670.

Gutschner, T., Hämmerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., et al. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. *73*, 1180–1189.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotech *28*, 503–510.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. Cell *154*, 240–251.

Haffter, P., Granato, M., Brand, M., Mullins, M.C., Hammerschmidt, M., Kane, D.A., Odenthal, J., Eeden, F.J. van, Jiang, Y.J., Heisenberg, C.P., et al. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development *123*, 1–36.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). BioMart Central Portal--unified access to biological data. Nucleic Acids Res. *37*, W23–27.

Hainer, S.J., Pruneski, J.A., Mitchell, R.D., Monteverde, R.M., and Martens, J.A. (2011). Intergenic transcription causes repression by directing nucleosome assembly. Genes Dev. *25*, 29–40.

Hamatani, T., Carter, M.G., Sharov, A.A., and Ko, M.S.H. (2004). Dynamics of global gene expression changes during mouse preimplantation development. Dev. Cell *6*, 117–131.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic

Noncoding RNAs. PLoS Genet *9*, e1003569.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Harvey, S.A., Sealy, I., Kettleborough, R., Fenyes, F., White, R., Stemple, D., and Smith, J.C. (2013). Identification of the zebrafish maternal and paternal transcriptomes. Dev. Camb. Engl. *140*, 2703–2710.

Al-Hasani, K., Vadolas, J., Knaupp, A.S., Wardan, H., Voullaire, L., Williamson, R., and Ioannou, P.A. (2005). A 191-kb genomic fragment containing the human alpha-globin locus can rescue alpha-thalassemic mice. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *16*, 847–853.

Hassan, M.A., Melo, M.B., Haas, B., Jensen, K.D.C., and Saeij, J.P.J. (2012). De novo reconstruction of the Toxoplasma gondii transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. BMC Genomics *13*, 696.

Haumaitre, C., Barbacci, E., Jenny, M., Ott, M.O., Gradwohl, G., and Cereghini, S. (2005). Lack of TCF2/vHNF1 in mice leads to pancreas agenesis. Proc. Natl. Acad. Sci. U. S. A. *102*, 1490–1495.

Hawkins, P.G., and Morris, K.V. (2010). Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. Transcription *1*, 165–175.

He, S., Liu, S., and Zhu, H. (2011a). The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol. Biol. *11*, 102.

He, S., Liu, S., and Zhu, H. (2011b). The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol. Biol. *11*, 102.

He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. (2008). The antisense transcriptomes of human cells. Science *322*, 1855–1857.

Heo, J.B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science *331*, 76–79.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Holmes, D.S., Mayfield, J.E., Sander, G., and Bonner, J. (1972). Chromosomal RNA: its properties. Science *177*, 72–74.

Hosack, D.A., Dennis, G., Jr, Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J.M., Frankish, A., Aken, B.L., Hourlier, T., et al. (2012). Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. Genome Res. *22*, 1698–1710.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome

sequence and its relationship to the human genome. Nature *496*, 498–503.

Huang, X., and Saint-Jeannet, J.-P. (2004). Induction of the neural crest and the opportunities of life on the edge. Dev. Biol. *275*, 1–11.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Hubbard, S.J., Grafham, D.V., Beattie, K.J., Overton, I.M., McLaren, S.R., Croning, M.D.R., Boardman, P.E., Bonfield, J.K., Burnside, J., Davies, R.M., et al. (2005). Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. Genome Res. *15*, 174–183.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M.K., Dibrov, A.A., Blanchard, A.A.A., Wang, X., Deng, G., Myal, Y., et al. (2006). Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. DNA Cell Biol. *25*, 418–428.

Hui, L., Ji, C., Hui, B., Lv, T., Ha, X., Yang, J., and Cai, W. (2009). The oncoprotein LMO3 interacts with calcium- and integrin-binding protein CIB. Brain Res. *1265*, 24–29.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., et al. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat. Genet. *43*, 621–629.

Hüttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet. TIG *21*, 289–297.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Inoue, T., Terada, K., Furukawa, A., Koike, C., Tamaki, Y., Araie, M., and Furukawa, T. (2006). Cloning and characterization of mr-s, a novel SAM domain protein, predominantly expressed in retinal photoreceptor cells. BMC Dev. Biol. *6*, 15.

Irimia, M., Tena, J.J., Alexis, M., Fernandez-Miñan, A., Maeso, I., Bogdanović, O., Calle-Mustienes, E.D.L., Roy, S.W., Gomez-Skarmeta, J.L., and Fraser, H.B. (2012). Extensive Conservation of Ancient Microsynteny Across Metazoans Due to Cis-Regulatory Constraints. Genome Res. *22*, 2356–2367.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature *431*, 946–957.

Jeon, Y., and Lee, J.T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119–133.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031–8041.

Ji, P., Liu, G., Xu, J., Wang, X., Li, J., Zhao, Z., Zhang, X., Zhang, Y., Xu, P., and Sun, X. (2012). Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. PloS One *7*, e35152.

Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA N. Y. N *16*, 1478–1487.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. TIG *21*, 93–102.

Kague, E., Gallagher, M., Burke, S., Parsons, M., Franz-Odendaal, T., and Fisher, S. (2012). Skeletogenic Fate of Zebrafish Cranial and Trunk Neural Crest. PLoS ONE *7*.

Kahn, S.E., Hull, R.L., and Utzschneider, K.M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature *444*, 840–846.

Kandalepas, P.C., and Vassar, R. (2012). Identification and biology of β-secretase. J. Neurochem. *120 Suppl 1*, 55–61.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916–919.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007a). Genome-wide transcription and the implications for genomic organization. Nat. Rev. Genet. *8*, 413–423.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007b). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C.P., Sorensen, P.H.B., Reaman, G., Milos, P., Arceci, R.J., Thompson, J.F., et al. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA. BMC Biol. *8*, 149.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet *9*, e1003470.

Karaskov, E., Scott, C., Zhang, L., Teodoro, T., Ravazzola, M., and Volchuk, A. (2006). Chronic Palmitate But Not Oleate Exposure Induces Endoplasmic Reticulum Stress, Which May Contribute to INS-1 Pancreatic β-Cell Apoptosis. Endocrinology *147*, 3398–3407.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the

mammalian transcriptome. Science *309*, 1564–1566.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. (2001). Functional annotation of a full-length mouse cDNA collection. Nature *409*, 685–690.

Kawashima, H., Takano, H., Sugita, S., Takahara, Y., Sugimura, K., and Nakatani, T. (2003). A novel steroid receptor co-activator protein (SRAP) as an alternative form of steroid receptor RNA-activator gene: expression in prostate cancer cells and enhancement of androgen receptor activity. Biochem. J. *369*, 163–171.

Ke, J., Xu, H.E., and Williams, B.O. (2013). Lipid modification in Wnt structure and function. Curr. Opin. Lipidol. *24*, 129–133.

Kelley, D.R., and Rinn, J.L. (2012). Transposable elements reveal a stem cell specific class of long noncoding RNAs. Genome Biol. *13*, R107.

Keniry, A., Oxley, D., Monnier, P., Kyba, M., Dandolo, L., Smits, G., and Reik, W. (2012). The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. Nat. Cell Biol. *14*, 659–665.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kesavadev, J.D., Short, K.R., and Nair, K.S. (2003). Diabetes in old age: an emerging epidemic. J. Assoc. Physicians India *51*, 1083–1094.

Kettleborough, R.N.W., Busch-Nentwich, E.M., Harvey, S.A., Dooley, C.M., de Bruijn, E., van Eeden, F., Sealy, I., White, R.J., Herd, C., Nijman, I.J., et al. (2013). A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature *496*, 494–497.

Khachane, A.N., and Harrison, P.M. (2010). Mining mammalian transcript data for functional long non-coding RNAs. PloS One *5*, e10316.

Khaitovich, P., Kelso, J., Franz, H., Visagie, J., Giger, T., Joerchel, S., Petzold, E., Green, R.E., Lachmann, M., and Pääbo, S. (2006). Functionality of intergenic transcription: an evolutionary comparison. PLoS Genet. *2*, e171.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D.R., Thomas, K., Presser, A., Bernstein, B.E., Oudenaarden, A. van, et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. *106*, 11667–11672.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007a). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. *17*, 545–555.

Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B., and Becker, T.S. (2007b). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. Genome Biol. *8*, S4.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013a). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kim, H.J., Sumanas, S., Palencia-Desai, S., Dong, Y., Chen, J.-N., and Lin, S. (2006). Genetic Analysis of Early Endocrine Pancreas Formation in Zebrafish. Mol. Endocrinol. *20*, 194–203.

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., Kim, S., and Safe, S. (2013b). HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene *32*, 1616–1625.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci. Signal. *3*, ra8.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database J. Biol. Databases Curation *2011*.

Kljuic, A., Bazzi, H., Sundberg, J.P., Martinez-Mir, A., O'Shaughnessy, R., Mahoney, M.G., Levy, M., Montagutelli, X., Ahmad, W., Aita, V.M., et al. (2003). Desmoglein 4 in hair follicle differentiation and epidermal adhesion: evidence from inherited hypotrichosis and acquired pemphigus vulgaris. Cell *113*, 249–260.

Kocabas, A.M., Crosby, J., Ross, P.J., Otu, H.H., Beyhan, Z., Can, H., Tam, W.-L., Rosa, G.J.M., Halgren, R.G., Lim, B., et al. (2006). The transcriptome of human oocytes. Proc. Natl. Acad. Sci. U. S. A. *103*, 14027–14032.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3*, 211–222.

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. Cancer Res. *71*, 6320–6326.

Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R., and Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. Epigenetics Chromatin *5*, 1.

Kondo, M., and Akasaka, K. (2012). Current Status of Echinoderm Genome Analysis - What do we Know? Curr. Genomics *13*, 134–143.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science *329*, 336–339.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345–349.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P., and Pauler, F.M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC Biol. *11*, 59.

Korostowski, L., Raval, A., Breuer, G., and Engel, N. (2011). Enhancer-driven chromatin

interactions during development promote escape from silencing by a long non-coding RNA. Epigenetics Chromatin *4*, 21.

Koski, L.B., Gray, M.W., Lang, B.F., and Burger, G. (2005). AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics *6*, 151.

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. Mol. Cell *45*, 447–458.

Kowalski, J.R., Dahlberg, C.L., and Juo, P. (2011). The deubiquitinating enzyme USP-46 negatively regulates the degradation of glutamate receptors to control their abundance in the ventral nerve cord of Caenorhabditis elegans. J. Neurosci. Off. J. Soc. Neurosci. *31*, 1341–1354.

Kraeussling, M., Wagner, T.U., and Schartl, M. (2011). Highly asynchronous and asymmetric cleavage divisions accompany early transcriptional activity in pre-blastula medaka embryos. PloS One *6*, e21741.

Kretz, M., Webster, D.E., Flockhart, R.J., Lee, C.S., Zehnder, A., Lopez-Pajares, V., Qu, K., Zheng, G.X.Y., Chow, J., Kim, G.E., et al. (2012). Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev. *26*, 338–343.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLoS Genet. *8*, e1002841.

Ladoukakis, E., Pereira, V., Magny, E.G., Eyre-Walker, A., and Couso, J.P. (2011). Hundreds of putatively functional small open reading frames in Drosophila. Genome Biol. *12*, R118.

Van de Lagemaat, L.N., Landry, J.-R., Mager, D.L., and Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. TIG *19*, 530–536.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *000*, 1.

Lam, S.H., and Gong, Z. (2006). Modeling liver cancer using zebrafish: a comparative oncogenomics approach. Cell Cycle Georget. Tex *5*, 573–577.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U., Wong, J., Tsai, S.Y., Tsai, M.J., and O'Malley, B.W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell *97*, 17–27.

Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R.,

Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., et al. (2012). Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. Science *338*, 1469–1472.

Laurent, G.S., Shtokalo, D., Dong, B., Tackett, M.R., Fan, X., Lazorthes, S., Nicolas, E., Sang, N., Triche, T.J., McCaffrey, T.A., et al. (2013). VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol. *14*, R73.

Lee, J.T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. *23*, 1831–1842.

Lee, J.T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nat. Rev. Mol. Cell Biol. *12*, 815–826.

Lee, J.T. (2012). Epigenetic Regulation by Long Noncoding RNAs. Science *338*, 1435–1439.

Lee, H.Y., Jung, H., Jang, I.H., Suh, P.-G., and Ryu, S.H. (2008a). Cdk5 phosphorylates PLD2 to mediate EGF-dependent insulin secretion. Cell. Signal. *20*, 1787–1794.

Lee, Y.-N., Gao, Y., and Wang, H.-Y. (2008b). Differential mediation of the Wnt canonical pathway by mammalian Dishevelleds-1, -2, and -3. Cell. Signal. *20*, 443–452.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012a). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Li, T., Wang, S., Wu, R., Zhou, X., Zhu, D., and Zhang, Y. (2012b). Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics *99*, 292–298.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013a). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Li, X., Stevens, P.D., Yang, H., Gulhati, P., Wang, W., Evers, B.M., and Gao, T. (2013b). The deubiquitination enzyme USP46 functions as a tumor suppressor by controlling PHLPP-dependent attenuation of Akt signaling in colon cancer. Oncogene *32*, 471–478.

Li, Z., Liu, M., Zhang, L., Zhang, W., Gao, G., Zhu, Z., Wei, L., Fan, Q., and Long, M. (2009). Detection of intergenic non-coding RNAs expressed in the main developmental stages in Drosophila melanogaster. Nucleic Acids Res. *37*, 4308–4314.

Liang, Q.L., Mo, Z., Li, X.F., Wang, X.X., and Li, R.M. (2013). Pdx1 protein induces human embryonic stem cells into the pancreatic endocrine lineage. Cell Biol. Int. *37*, 2–10.

Licastro, D., Gennarino, V.A., Petrera, F., Sanges, R., Banfi, S., and Stupka, E. (2010). Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. BMC Genomics *11*, 151.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lim, J.E., Hong, K.-W., Jin, H.-S., Kim, Y.S., Park, H.K., and Oh, B. (2010). Type 2 diabetes genetic association database manually curated for the study design and odds

ratio. BMC Med. Inform. Decis. Mak. *10*, 76.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773.

Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., and Lachman, H.M. (2011a). RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PloS One *6*, e23356.

Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., and Lachman, H.M. (2011b). RNA-Seq of Human Neurons Derived from iPS Cells Reveals Candidate Long Non-Coding RNAs Involved in Neurogenesis and Neuropsychiatric Disorders. PLoS ONE *6*.

Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St Pierre, S.E., et al. (2007). Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res. *17*, 1823–1836.

Lin, M.F., Deoras, A.N., Rasmussen, M.D., and Kellis, M. (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. PLoS Comput. Biol. *4*, e1000067.

Lin, M.F., Jungreis, I., and Kellis, M. (2011c). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275 –i282.

Lin, Q., Schwarz, J., Bucana, C., and Olson, E.N. (1997). Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. Science *276*, 1404–1407.

Lipovich, L., Tarca, A.L., Cai, J., Jia, H., Chugani, H.T., Sterner, K.N., Grossman, L.I., Uddin, M., Hof, P.R., Sherwood, C.C., et al. (2013). Developmental Changes in the Transcriptome of Human Cerebral Cortex Tissue: Long Noncoding RNA Transcripts. Cereb. Cortex N. Y. N 1991.

Liu, S., and Leach, S.D. (2011). Zebrafish models for cancer. Annu. Rev. Pathol. *6*, 71–93.

Liu, H., Wang, T., Wang, J., Quan, F., and Zhang, Y. (2013). Characterization of liaoning cashmere goat transcriptome: sequencing, de novo assembly, functional annotation and comparative analysis. PloS One *8*, e77062.

Liu, J., Gough, J., and Rost, B. (2006). Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. PLoS Genet *2*, e29.

Liu, Y., Fallon, L., Lashuel, H.A., Liu, Z., and Lansbury, P.T., Jr (2002). The UCH-L1 gene encodes two opposing enzymatic activities that affect alpha-synuclein degradation and Parkinson's disease susceptibility. Cell *111*, 209–218.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. *40*, W622–627.

Lu, F.-I., Thisse, C., and Thisse, B. (2011). Identification and mechanism of regulation of the zebrafish dorsal determinant. Proc. Natl. Acad. Sci. *108*, 15876–15880.

Lu, T., Zhu, C., Lu, G., Guo, Y., Zhou, Y., Zhang, Z., Zhao, Y., Li, W., Lu, Y., Tang, W., et al.

(2012). Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice. BMC Genomics *13*, 721.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., et al. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet. *2*, e62.

Maher, B. (2012). ENCODE: The human encyclopaedia. Nature *489*, 46–48.

Malik, S., and Roeder, R.G. (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat. Rev. Genet. *11*, 761–772.

Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., Langford, C.F., Ost, T.W.B., Collins, J.E., and Turner, D.J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods *7*, 130–132.

Managadze, D., Lobkovsky, A.E., Wolf, Y.I., Shabalina, S.A., Rogozin, I.B., and Koonin, E.V. (2013). The Vast, Conserved Mammalian lincRNome. PLoS Comput Biol *9*, e1002917.

Manni, I., Caretti, G., Artuso, S., Gurtner, A., Emiliozzi, V., Sacchi, A., Mantovani, R., and Piaggio, G. (2008). Posttranslational regulation of NF-YA modulates NF-Y transcriptional activity. Mol. Biol. Cell *19*, 5203–5213.

Mao, Y.S., Zhang, B., and Spector, D.L. (2011). Biogenesis and function of nuclear bodies. Trends Genet. TIG *27*, 295–306.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., et al. (2012). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. *41*, D348–352.

Mariner, P.D., Walters, R.D., Espinoza, C.A., Drullinger, L.F., Wagner, S.D., Kugel, J.F., and Goodrich, J.A. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol. Cell *29*, 499–509.

Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. *10*, R124.

Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol. *14*, R131.

Mathavan, S., Lee, S.G.P., Mak, A., Miller, L.D., Murthy, K.R.K., Govindarajan, K.R., Tong, Y., Wu, Y.L., Lam, S.H., Yang, H., et al. (2005). Transcriptome Analysis of Zebrafish Embryogenesis Using Microarrays. PLoS Genet *1*, e29.

Matschinsky, F.M. (2002). Regulation of pancreatic beta-cell glucokinase: from basics to therapeutics. Diabetes *51 Suppl 3*, S394–404.

Matsuda, H., Parsons, M.J., and Leach, S.D. (2013). Aldh1-expressing endocrine progenitor cells regulate secondary islet formation in larval zebrafish pancreas. PloS One *8*, e74350.

Matsui, K., Nishizawa, M., Ozaki, T., Kimura, T., Hashimoto, I., Yamada, M., Kaibori, M., Kamiyama, Y., Ito, S., and Okumura, T. (2008). Natural antisense transcript stabilizes inducible nitric oxide synthase messenger RNA in rat hepatocytes. Hepatol. Baltim. Md *47*, 686–697.

Mattick, J.S. (2011). The central role of RNA in human development and cognition. FEBS Lett. *585*, 1600–1616.

Mayo, K.E., Cerelli, G.M., Rosenfeld, M.G., and Evans, R.M. (1985). Characterization of cDNA and genomic clones encoding the precursor to rat hypothalamic growth hormone-releasing factor. Nature *314*, 464–467.

McClelland, A.D., and Kantharidis, P. (2014). microRNA in the development of diabetic complications. Clin. Sci. Lond. Engl. 1979 *126*, 95–110.

McCutcheon, J.P., and Eddy, S.R. (2003). Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res. *31*, 4119–4128.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods *6*, S13–S20.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. U. S. A. *105*, 716–721.

Mercer, T.R., Qureshi, I.A., Gokhan, S., Dinger, M.E., Li, G., Mattick, J.S., and Mehler, M.F. (2010). Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. BMC Neurosci. *11*, 14.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2011). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. *30*, 99–104.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2012). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *40*, D64–69.

Minoux, M., Antonarakis, G.S., Kmita, M., Duboule, D., and Rijli, F.M. (2009). Rostral and caudal pharyngeal arches share a common neural crest ground pattern. Development *136*, 637–645.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O., and Lai, E.C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. *23*, 812–825.

Modarresi, F., Faghihi, M.A., Lopez-Toledano, M.A., Fatemi, R.P., Magistri, M., Brothers, S.P., van der Brug, M.P., and Wahlestedt, C. (2012). Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. Nat. Biotechnol. *30*, 453–459.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. Cell *147*, 1132–1145.

Morachis, J.M., Murawsky, C.M., and Emerson, B.M. (2010). Regulation of the p53 transcriptional response by structurally diverse core promoters. Genes Dev. *24*, 135–147.

Morán, I., Akerman, İ., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J., Rodríguez-Seguí, S., et al. (2012). Human β Cell Transcriptome Analysis Uncovers lncRNAs That Are Tissue-Specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes. Cell Metab. *16*, 435–448.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques *45*, 81–94.

Morishita, H., and Yagi, T. (2007). Protocadherin family: diversity, structure, and function. Curr. Opin. Cell Biol. *19*, 584–592.

Morris, K.V., and Vogt, P.K. (2010). Long antisense non-coding RNAs and their role in transcription and oncogenesis. Cell Cycle *9*, 2544–2547.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Moss, J.B., Koustubhan, P., Greenman, M., Parsons, M.J., Walter, I., and Moss, L.G. (2009). Regeneration of the Pancreas in Adult Zebrafish. Diabetes *58*, 1844–1851.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Mowry, K.L., and Cote, C.A. (1999). RNA sorting in Xenopus oocytes and embryos. FASEB J. *13*, 435–445.

Mujoo, K., Krumenacker, J.S., and Murad, F. (2011). Nitric oxide-cyclic GMP signaling in stem cell differentiation. Free Radic. Biol. Med. *51*, 2150–2157.

Mustafi, D., Kevany, B.M., Bai, X., Maeda, T., Sears, J.E., Khalil, A.M., and Palczewski, K. (2013). Evolutionarily conserved long intergenic non-coding RNAs in the eye. Hum. Mol. Genet.

Nakagawa, S., Naganuma, T., Shioi, G., and Hirose, T. (2011). Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. J. Cell Biol. *193*, 31–39.

Nakamura, T., Nishizawa, T., Hagiya, M., Seki, T., Shimonishi, M., Sugimura, A., Tashiro, K., and Shimizu, S. (1989). Molecular cloning and expression of human hepatocyte growth factor. Nature *342*, 440–443.

Nam, J.-W., and Bartel, D. (2012). Long non-coding RNAs in C. elegans. Genome Res. *22*, 2529–2540.

Van Name, M., and Santoro, N. (2013). Type 2 diabetes mellitus in pediatrics: a new challenge. World J. Pediatr. WJP *9*, 293–299.

Nathan, C., and Xie, Q.W. (1994). Nitric oxide synthases: roles, tolls, and controls. Cell *78*, 915–918.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M.M., Sheng, Y., Abdelhamid, R.F., Anand, S., et al. (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate

embryogenesis. Genome Res. *23*, 1938–1950.

Newport, J., and Kirschner, M. (1982). A major developmental transition in early xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage. Cell *30*, 675–686.

Ni, T., Tu, K., Wang, Z., Song, S., Wu, H., Xie, B., Scott, K.C., Grewal, S.I., Gao, Y., and Zhu, J. (2010). The prevalence and regulation of antisense transcripts in Schizosaccharomyces pombe. PloS One *5*, e15271.

Niinuma, T., Suzuki, H., Nojima, M., Nosho, K., Yamamoto, H., Takamaru, H., Yamamoto, E., Maruyama, R., Nobuoka, T., Miyazaki, Y., et al. (2012). Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. Cancer Res. *72*, 1126–1136.

Nishimoto, Y., Nakagawa, S., Hirose, T., Okano, H.J., Takao, M., Shibata, S., Suyama, S., Kuwako, K.-I., Imai, T., Murayama, S., et al. (2013). The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis. Mol. Brain *6*, 31.

Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., Janssen-Megens, E.M., Hubner, N.C., Schulz, H., de Vries, I., Ding, L., Huebner, N., Mann, M., et al. (2011). RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. PloS One *6*, e19470.

Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. (2012). Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure? BioArchitecture *2*, 189–199.

Odawara, J., Harada, A., Yoshimi, T., Maehara, K., Tachibana, T., Okada, S., Akashi, K., and Ohkawa, Y. (2011). The classification of mRNA expression levels by the phosphorylation state of RNAPII CTD based on a combined genome-wide approach. BMC Genomics *12*, 516.

Ogawa, Y., Sun, B.K., and Lee, J.T. (2008). Intersection of the RNA interference and X-inactivation pathways. Science *320*, 1336–1341.

Ohhata, T., Senner, C.E., Hemberger, M., and Wutz, A. (2011). Lineage-specific function of the noncoding Tsix RNA for Xist repression and Xi reactivation in mice. Genes Dev. *25*, 1702–1715.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

Olety, B., Wälte, M., Honnert, U., Schillers, H., and Bähler, M. (2010). Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity. FEBS Lett. *584*, 493–499.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010a). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010b). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A.,

Hayashi, K., Sato, H., Nagai, K., et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat. Genet. *36*, 40–45.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat. Rev. Drug Discov. *5*, 993–996.

De Palma, C., and Clementi, E. (2012). Nitric oxide in myogenesis and therapeutic muscle repair. Mol. Neurobiol. *46*, 682–692.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. TIG *22*, 1–5.

Pang, K.C., Dinger, M.E., Mercer, T.R., Malquori, L., Grimmond, S.M., Chen, W., and Mattick, J.S. (2009). Genome-wide identification of long noncoding RNAs in CD8+ T cells. J. Immunol. Baltim. Md 1950 *182*, 7738–7748.

Paranjpe, S.S., Jacobi, U.G., Heeringen, S.J. van, and Veenstra, G.J.C. (2013). A genome-wide survey of maternal and embryonic transcripts during Xenopus tropicalis development. BMC Genomics *14*, 762.

Park, S.W., Davison, J.M., Rhee, J., Hruban, R.H., Maitra, A., and Leach, S.D. (2008). Oncogenic KRAS induces progenitor cell expansion and malignant transformation in zebrafish exocrine pancreas. Gastroenterology *134*, 2080–2090.

Pashos, E., Park, J.T., Leach, S., and Fisher, S. (2013). Distinct enhancers of ptf1a mediate specification and expansion of ventral pancreas in zebrafish. Dev. Biol. *381*, 471–481.

Passalacqua, K.D., Varadarajan, A., Weist, C., Ondov, B.D., Byrd, B., Read, T.D., and Bergman, N.H. (2012). Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in Bacillus anthracis. PloS One *7*, e43350.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., et al. (2011a). Systematic identification of long non-coding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pauli, A., Rinn, J.L., and Schier, A.F. (2011b). Non-coding RNAs as regulators of embryogenesis. Nat. Rev. Genet. *12*, 136–149.

Peer, Y.V. de (2004). Tetraodon genome confirms Takifugu findings: most fish are ancient polyploids. Genome Biol. *5*, 250.

Pelegri, F. (2003). Maternal factors in zebrafish development. Dev. Dyn. Off. Publ. Am. Assoc. Anat. *228*, 535–554.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Pennisi, E. (2012). ENCODE Project Writes Eulogy for Junk DNA. Science *337*, 1159–1161.

Pettitt, D.J., Talton, J., Dagraphics6belea, D., Divers, J., Imperatore, G., Lawrence, J.M., Liese, A.D., Linder, B., Mayer-Davis, E.J., Pihoker, C., et al. (2013). Prevalence of Diabetes Mellitus in U.S. Youth in 2009: The SEARCH for Diabetes in Youth Study. Diabetes Care.

Petzold, A.M., Balciunas, D., Sivasubbu, S., Clark, K.J., Bedell, V.M., Westcot, S.E., Myers, S.R., Moulder, G.L., Thomas, M.J., and Ekker, S.C. (2009). Nicotine response genetics in the zebrafish. Proc. Natl. Acad. Sci. *106*, 18662–18667.

Philipp, E.E.R., Kraemer, L., Mountfort, D., Schilhabel, M., Schreiber, S., and Rosenstiel, P. (2012). The Transcriptome Analysis and Comparison Explorer—T-ACE: A Platform-Independent, Graphical Tool to Process Large RNAseq Datasets of Non-Model Organisms. Bioinformatics *28*, 777–783.

Pierpont, M.E., and Yunis, J.J. (1977). Localization of chromosomal RNA in human G-banded metaphase chromosomes. Exp. Cell Res. *106*, 303–308.

Pisharath, H., Rhee, J.M., Swanson, M.A., Leach, S.D., and Parsons, M.J. (2007). Targeted ablation of beta cells in the embryonic zebrafish pancreas using E. coli nitroreductase. Mech. Dev. *124*, 218–229.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006a). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006b). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006c). Forces shaping the fastest evolving regions in the human genome. PLoS Genet. *2*, e168.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Ponjavic, J., and Ponting, C.P. (2007). The long and the short of RNA maps. BioEssays News Rev. Mol. Cell. Dev. Biol. *29*, 1077–1080.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.

Pontier, D.B., and Gribnau, J. (2011). Xist regulation and function eXplored. Hum. Genet.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M.J., Stabenau, A., Storey, R., and Clamp, M. (2004). The Ensembl Analysis Pipeline. Genome Res. *14*, 934–941.

Prasad, T., and Weiner, J.A. (2011). Direct and Indirect Regulation of Spinal Cord Ia Afferent Terminal Formation by the γ-Protocadherins. Front. Mol. Neurosci. *4*, 54.

Pueyo, J.I., and Couso, J.P. (2011). Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. Dev. Biol. *355*, 183–193.

Pushkarev, D., Neff, N.F., and Quake, S.R. (2009). Single-molecule sequencing of an individual human genome. Nat. Biotechnol. *27*, 847–850.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A.,

Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86–94.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064–1071.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842.

Qureshi, I.A., and Mehler, M.F. (2012). Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat. Rev. Neurosci. *13*, 528–541.

Qureshi, I.A., Mattick, J.S., and Mehler, M.F. (2010). Long non-coding RNAs in nervous system function and disease. Brain Res. *1338*, 20–35.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., et al. (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628–1639.

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K.-Y., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., and Lim, D.A. (2013). Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. Cell Stem Cell *12*, 12–15.

Reddy, R., Henning, D., and Busch, H. (1979). Nucleotide sequence of nucleolar U3B RNA. J. Biol. Chem. *254*, 11097–11105.

Rein, A. (1971). The small molecular weight monodisperse nuclear RNA's in mitotic cells. Biochim. Biophys. Acta *232*, 306–313.

Rinn, J.L., and Chang, H.Y. (2012). Genome Regulation by Long Noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M., et al. (2003). The transcriptional activity of human Chromosome 22. Genes Dev. *17*, 529–540.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Robinson, M.D., and Smyth, G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinforma. Oxf. Engl. *23*, 2881–2887.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140.

Da Rocha, S.T., Edwards, C.A., Ito, M., Ogata, T., and Ferguson-Smith, A.C. (2008). Genomic imprinting at the mammalian Dlk1-Dio3 domain. Trends Genet. TIG *24*, 306–316.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., et al. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat.

Genet. *25*, 235–238.

Rokyta, D.R., Lemmon, A.R., Margres, M.J., and Aronow, K. (2012). The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). BMC Genomics *13*, 312.

Sadamoto, H., Takahashi, H., Okada, T., Kenmoku, H., Toyota, M., and Asakawa, Y. (2012). De novo sequencing and transcriptome analysis of the central nervous system of mollusc Lymnaea stagnalis by deep RNA sequencing. PloS One *7*, e42546.

Sado, T., Hoki, Y., and Sasaki, H. (2005). Tsix silences Xist through modification of chromatin structure. Dev. Cell *9*, 159–165.

Sakuraba, Y., Kimura, T., Masuya, H., Noguchi, H., Sezutsu, H., Takahasi, K.R., Toyoda, A., Fukumura, R., Murata, T., Sakaki, Y., et al. (2008). Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *19*, 703–712.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Sandelin, A., Bailey, P., Bruce, S., Engström, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

Sanges, R., Hadzhiev, Y., Gueroult-Bellone, M., Roure, A., Ferg, M., Meola, N., Amore, G., Basu, S., Brown, E.R., De Simone, M., et al. (2013). Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. Nucleic Acids Res. *41*, 3600–3618.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Santini, S., Boore, J.L., and Meyer, A. (2003). Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res. *13*, 1111–1122.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schloss, P.D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. *6*, e1000844.

Schmid, R., and Blaxter, M.L. (2008). annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics *9*, 180.

Schmidt, L.H., Spieker, T., Koschmieder, S., Schäffers, S., Humberg, J., Jungen, D., Bulk, E., Hascher, A., Wittmer, D., Marra, A., et al. (2011). The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer *6*, 1984–1992.

Schmitz, K.-M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev. *24*, 2264–2269.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long

non-coding RNA hotair in mouse and human. PLoS Genet. *7*, e1002071.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat. Methods *5*, 16–18.

Seemann, S.E., Gilchrist, M.J., Hofacker, I.L., Stadler, P.F., and Gorodkin, J. (2007). Detection of RNA structures in porcine EST data and related mammals. BMC Genomics *8*, 316.

Seitz, H. (2009). Redefining microRNA targets. Curr. Biol. CB *19*, 870–873.

Seth, A., Stemple, D.L., and Barroso, I. (2013). The emerging use of zebrafish to model metabolic disease. Dis. Model. Mech. *6*, 1080–1088.

Shao, W.-J., Tao, L.-Y., Gao, C., Xie, J.-Y., and Zhao, R.-Q. (2008). Alterations in Methylation and Expression Levels of Imprinted Genes H19 and Igf2 in the Fetuses of Diabetic Mice. Comp. Med. *58*, 341–346.

Sheik Mohamed, J., Gaughwin, P.M., Lim, B., Robson, P., and Lipovich, L. (2010). Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. RNA N. Y. N *16*, 324–337.

Shen, H., McElhinny, A.S., Cao, Y., Gao, P., Liu, J., Bronson, R., Griffin, J.D., and Wu, L. (2006). The Notch coactivator, MAML1, functions as a novel coactivator for MEF2C-mediated transcription and is required for normal myogenesis. Genes Dev. *20*, 675–688.

Shen, S.J., Daimon, M., Wang, C.Y., Jansen, M., and Ilan, J. (1988). Isolation of an insulin-like growth factor II cDNA with a unique 5' untranslated region from human placenta. Proc. Natl. Acad. Sci. U. S. A. *85*, 1947–1951.

Shimizu, K., Adachi, J., and Muraoka, Y. (2006). ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. J. Bioinform. Comput. Biol. *4*, 649–664.

Shin, J.T., Priest, J.R., Ovcharenko, I., Ronco, A., Moore, R.K., Burns, C.G., and MacRae, C.A. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res. *33*, 5437–5445.

Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc. Natl. Acad. Sci. U. S. A. *109*, 1347–1352.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. *110*, 2876–2881.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinforma. Oxf. Engl. *21*, 3940–3941.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. *9*, 657–663.

Smith, C.M., and Steitz, J.A. (1998). Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. Mol. Cell. Biol. *18*, 6897–6909.

Smyk, M., Szafranski, P., Startek, M., Gambin, A., and Stankiewicz, P. (2013). Chromosome conformation capture-on-chip analysis of long-range cis-interactions of the SOX9 promoter. Chromosome Res. *21*, 781–788.

Sone, M., Hayashi, T., Tarui, H., Agata, K., Takeichi, M., and Nakagawa, S. (2007). The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J. Cell Sci. *120*, 2498–2506.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The Trichoplax genome and the nature of placozoans. Nature *454*, 955–960.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. Nature *466*, 720–726.

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., and Birney, E. (2004). The Ensembl Core Software Libraries. Genome Res. *14*, 929–933.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res. *12*, 1611–1618.

Stamatoyannopoulos, J.A. (2012). What does our genome encode? Genome Res. *22*, 1602–1611.

St Laurent, G., 3rd, and Wahlestedt, C. (2007). Noncoding RNAs: couplers of analog and digital information in nervous system function? Trends Neurosci. *30*, 612–621.

Stöhr, R., and Federici, M. (2013). Insulin resistance and atherosclerosis: convergence between metabolic pathways and inflammatory nodes. Biochem. J. *454*, 1–11.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Su, H., Marcheva, B., Meng, S., Liang, F.A., Kohsaka, A., Kobayashi, Y., Xu, A.W., Bass, J., and Wang, X. (2010). Gamma-protocadherins regulate the functional integrity of hypothalamic feeding circuitry in mice. Dev. Biol. *339*, 38–50.

Sun, B.K., Deaton, A.M., and Lee, J.T. (2006). A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. Mol. Cell *21*, 617–628.

Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., and Sun, H. (2013a). iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics *14*, S7.

Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., and Lee, J.T. (2013b). Jpx RNA activates Xist by evicting CTCF. Cell *153*, 1537–1551.

Sun, Z., Amsterdam, A., Pazour, G.J., Cole, D.G., Miller, M.S., and Hopkins, N. (2004). A genetic screen in zebrafish identifies cilia genes as a principal cause of cystic kidney. Dev. Camb. Engl. *131*, 4085–4093.

Svoboda, P., and Flemr, M. (2010). The role of miRNAs and endogenous siRNAs in maternal-to-zygotic reprogramming and the establishment of pluripotency. EMBO Rep. *11*, 590–597.

Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033–3042.

Tajbakhsh, S., and Cossu, G. (1997). Establishing myogenic identity during somitogenesis. Curr. Opin. Genet. Dev. *7*, 634–641.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tani, H., and Torimura, M. (2013). Identification of short-lived long non-coding RNAs as surrogate indicators for chemical stress response. Biochem. Biophys. Res. Commun. *439*, 547–551.

Taylor, S.I. (1999). Deconstructing type 2 diabetes. Cell *97*, 9–12.

The ENCODE Project Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

The ENCODE Project Consortium, T.E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

The Gene Ontology Consortium (2012). Gene Ontology Annotations and Resources. Nucleic Acids Res. *41*, D530–D535.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2012). Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. Brief. Bioinform. *14*, 178–192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390–403.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Tiwari, A., Schuiki, I., Zhang, L., Allister, E.M., Wheeler, M.B., and Volchuk, A. (2013). SDF2L1 interacts with the ER-associated degradation machinery and retards the degradation of mutant proinsulin in pancreatic β-cells. J. Cell Sci. *126*, 1962–1968.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinforma. Oxf. Engl. *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell *39*, 925–938.

Tripathi, V., Shen, Z., Chakraborty, A., Giri, S., Freier, S.M., Wu, X., Zhang, Y., Gorospe, M., Prasanth, S.G., Lal, A., et al. (2013). Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet. *9*, e1003368.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M., and Nakagawa, S. (2011). Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. Genes Cells Devoted Mol. Cell. Mech. *16*, 479–490.

Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.A., and Davidson, E.H. (2012). Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res. *22*, 2079–2087.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell *147*, 1537–1550.

UniProt Consortium (2009). The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res. *37*, D169–174.

Vassar, R., Kovacs, D.M., Yan, R., and Wong, P.C. (2009). The beta-secretase enzyme BACE in health and Alzheimer's disease: regulation, cell biology, function, and therapeutic potential. J. Neurosci. Off. J. Soc. Neurosci. *29*, 12787–12794.

Vavouri, T., Walter, K., Gilks, W.R., Lehner, B., and Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. *8*, R15.

Venkatraman, A., He, X.C., Thorvaldsen, J.L., Sugimura, R., Perry, J.M., Tao, F., Zhao, M., Christenson, M.K., Sanchez, R., Yu, J.Y., et al. (2013). Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence. Nature *500*, 345–349.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., and Mestdagh, P. (2012). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res.

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. Nat. Rev. Drug Discov. *12*, 433–446.

Wallace, H.A.C., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J.A., Hughes, J., Wood, W.G., Higgs, D.R., and Smith, A.J.H. (2007). Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. Cell *128*, 197–209.

Wallace, K.N., Akhter, S., Smith, E.M., Lorent, K., and Pack, M. (2005). Intestinal growth and differentiation in zebrafish. Mech. Dev. *122*, 157–173.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K.-S. (2004). Mouse transcriptome: Neutral evolution of "non-coding" complementary DNAs. Nature *431*.

Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F., and Fan, Q. (2010). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. *38*, 5366–5383.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Ward, L.D., and Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science *337*, 1675–1678.

Washietl, S., Hofacker, I.L., and Stadler, P.F. (2005). Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U. S. A. *102*, 2454–2459.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539–543.

Watson, C.A., Hill, J.E., Graves, J.S., Wood, A.L., and Kilgore, K.H. (2009). Use of a novel induced spawning technique for the first reported captive spawning of Tetraodon nigroviridis. Mar. Genomics *2*, 143–146.

Weaver, C., and Kimelman, D. (2004). Move it or lose it: axis specification in Xenopus. Dev. Camb. Engl. *131*, 3491–3499.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

Wernersson, R. (2006). Virtual Ribosome--a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res. *34*, W385–388.

Wienholds, E., Eeden, F. van, Kosters, M., Mudde, J., Plasterk, R.H.A., and Cuppen, E. (2003). Efficient Target-Selected Mutagenesis in Zebrafish. Genome Res. *13*, 2700–2707.

Wilfinger, A., Arkhipova, V., and Meyer, D. (2013). Cell type and tissue specific function of islet genes in zebrafish pancreas development. Dev. Biol. *378*, 25–37.

Wood, S.H., Craig, T., Li, Y., Merry, B., and de Magalhães, J.P. (2012). Whole transcriptome sequencing of the aging rat brain reveals dynamic RNA changes in the dark matter of the genome. Age Dordr. Neth.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F.,

North, P., Callaway, H., Kelly, K., et al. (2005). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. PLoS Biol. *3*, e7.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. TIG *25*, 434–440.

Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat. Rev. Genet. *12*, 542–553.

Xu, X., Meiler, S.E., Zhong, T.P., Mohideen, M., Crossley, D.A., Burggren, W.W., and Fishman, M.C. (2002). Cardiomyopathy in zebrafish due to mutation in an alternatively spliced exon of titin. Nat. Genet. *30*, 205–209.

Xu, Z., Wei, G., Chepelev, I., Zhao, K., and Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. Nat. Struct. Mol. Biol. *18*, 372–378.

Yang, L., Lin, C., Jin, C., Yang, J.C., Tanasa, B., Li, W., Merkurjev, D., Ohgi, K.A., Meng, D., Zhang, J., et al. (2013a). lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature *500*, 598–602.

Yang, X., Song, J.H., Cheng, Y., Wu, W., Bhagat, T., Yu, Y., Abraham, J.M., Ibrahim, S., Ravich, W., Roland, B.C., et al. (2013b). Long non-coding RNA HNF1A-AS1 regulates proliferation and migration in oesophageal adenocarcinoma cells. Gut.

Yang, Y.H.C., Fox, J.E.M., Zhang, K.L., MacDonald, P.E., and Johnson, J.D. (2013c). Intraislet SLIT–ROBO signaling is required for beta-cell survival and potentiates insulin secretion. Proc. Natl. Acad. Sci. 201214312.

Yates, L.A., Norbury, C.J., and Gilbert, R.J.C. (2013). The long and short of microRNA. Cell *153*, 516–519.

Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., et al. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. *13*, R48.

Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol. Cell *47*, 648–655.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.-L., and Ponting, C.P. (2012). Identification and Properties of 1,119 Candidate LincRNA Loci in the Drosophila melanogaster Genome. Genome Biol. Evol. *4*, 427–442.

Zeng, D., Chen, X., Xie, D., Zhao, Y., Yang, C., Li, Y., Ma, N., Peng, M., Yang, Q., Liao, Z., et al. (2013). Transcriptome analysis of Pacific white shrimp (Litopenaeus vannamei) hepatopancreas in response to Taura syndrome Virus (TSV) experimental infection. PloS One *8*, e57515.

ZeRuth, G.T., Takeda, Y., and Jetten, A.M. (2013). The Krüppel-like protein Gli-similar 3 (Glis3) functions as a key regulator of insulin transcription. Mol. Endocrinol. Baltim. Md *27*, 1692–1705.

Zeyfang, A., and Bahrmann, A. (2013). [Diabetes in old age--risk by over- and undertreatment]. MMW Fortschr. Med. *155*, 56–58.

Zhang, X.H.-F., and Chasin, L.A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc. Natl. Acad. Sci. U. S. A. *103*, 13427–13432.

Zhang, X., Lian, Z., Padden, C., Gerstein, M.B., Rozowsky, J., Snyder, M., Gingeras, T.R., Kapranov, P., Weissman, S.M., and Newburger, P.E. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood *113*, 2526–2534.

Zhang, X., Mao, Y., Huang, Z. Qu, M., Chen, J., Ding, S., Hong, J., and Sun, T. (2012). Transcriptome Analysis of the Octopus vulgaris Central Nervous System. PLoS ONE *7*, e40320.

Zhang, Z., Theler, D., Kaminska, K.H., Hiller, M., de la Grange, P., Pudimat, R., Rafalska, I., Heinrich, B., Bujnicki, J.M., Allain, F.H.-T., et al. (2010). The YTH domain is a novel RNA binding domain. J. Biol. Chem. *285*, 14701–14710.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. *11*, R124.

Zhu, L., and Xu, P.-C. (2013). Downregulated LncRNA-ANCR promotes osteoblast differentiation by targeting EZH2 and regulating Runx2 expression. Biochem. Biophys. Res. Commun. *432*, 612–617.

Zschäbitz, A., Krahn, V., Schmidt, W., Gabius, H.J., Weiser, H., Biesalski, H.K., Kunt, T., Koepp, H., and Stofft, E. (1999). Expression patterns of complex glycoconjugates and endogenous lectins during fetal development of the viscerocranium. Ann. Anat. Anat. Anz. Off. Organ Anat. Ges. *181*, 117–121.

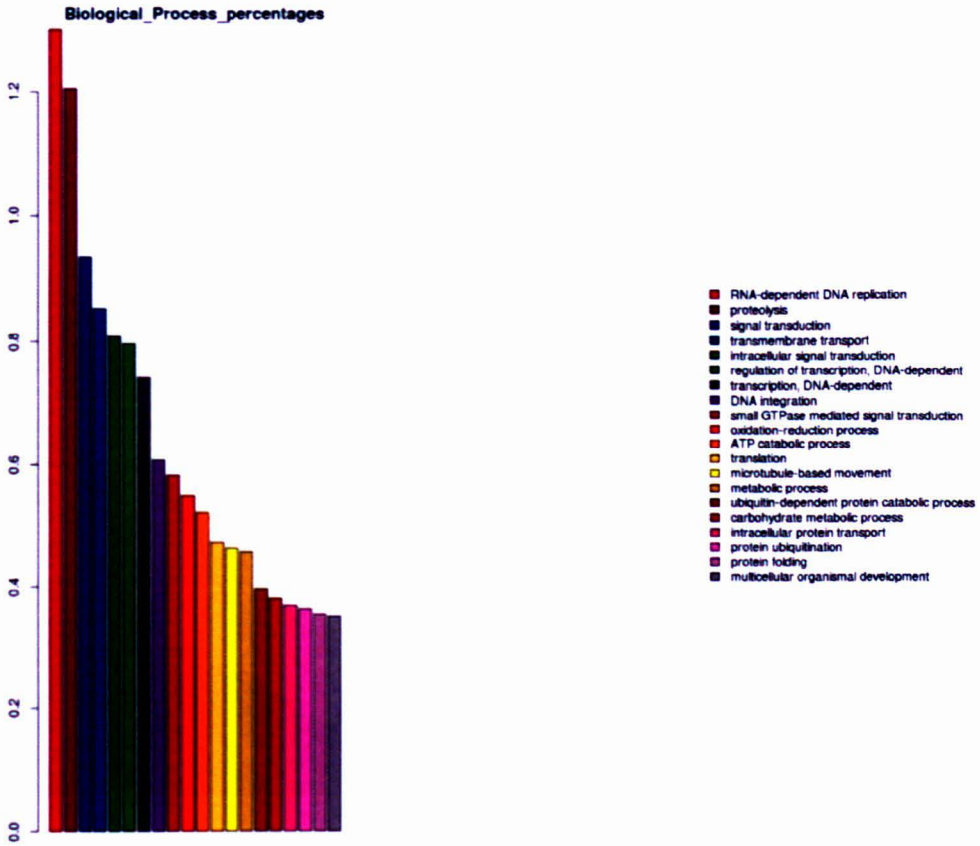
Annexure 1

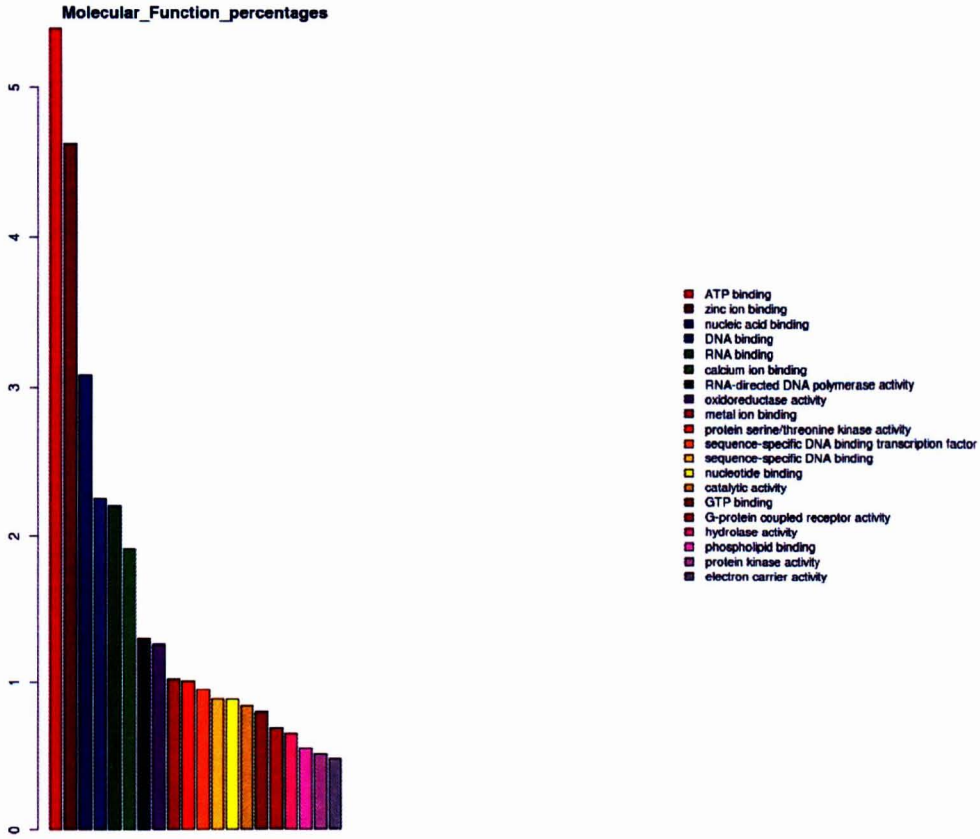
Annocript 2.0 – Results

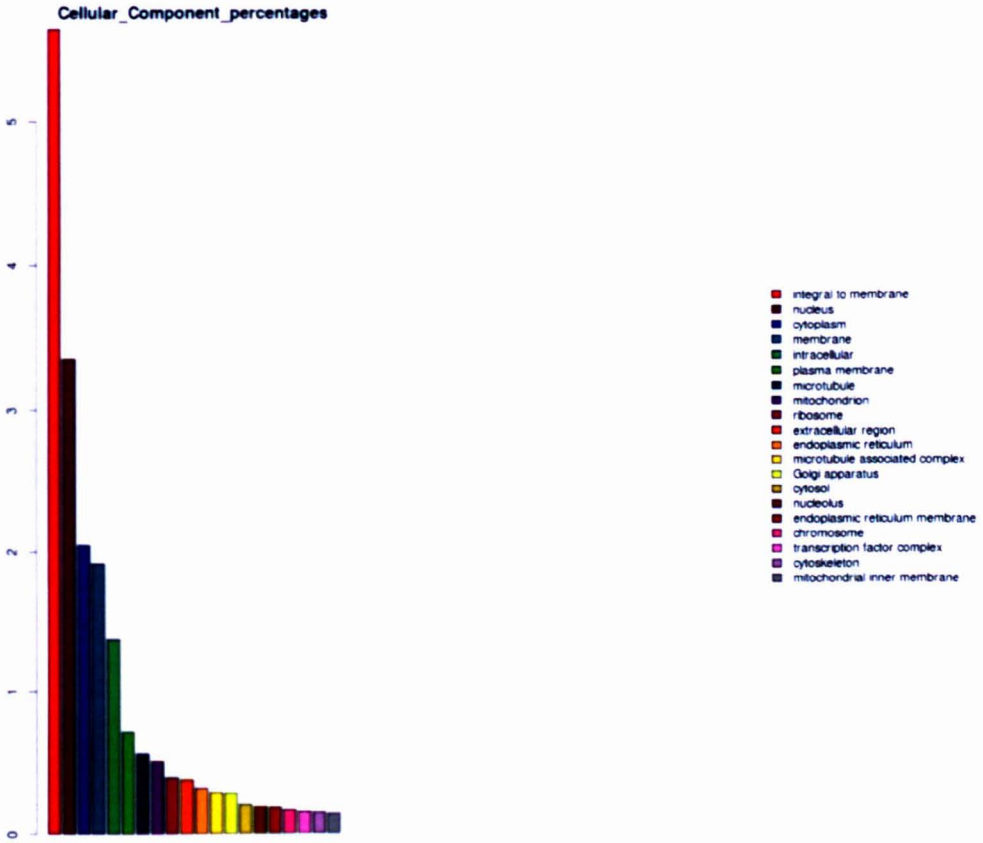
Statistics for transcriptome:

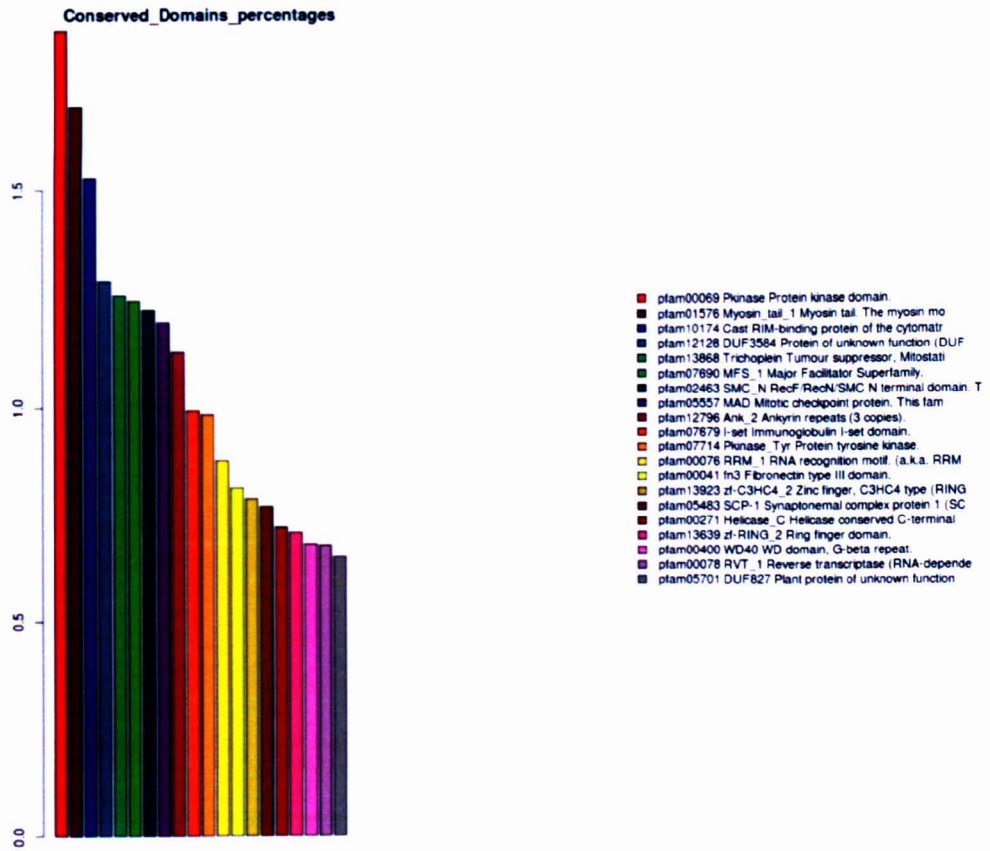
The file of sequences is
/home/francesco/ann_works/jobs/astropecten2/starfish_transcriptome_2013_10_filtered.fasta
The total number of sequences is 64388
The mean sequences length is 2125
The minimum and maximum sequences length are respectively 201 and 34228
Mean percentage of Adenine: 28.99 ;
Mean percentage of Guanine: 21.09 ;
Mean percentage of Thymine: 29.05 ;
Mean percentage of Cytosine: 20.86 ;
Mean percentage of N: 0.00 ;
Mean percentage of CG: 41.95
Number of annotated sequences: 32783
Sequences in agreement with strand info: 29564
Number of non coding sequences: 3133

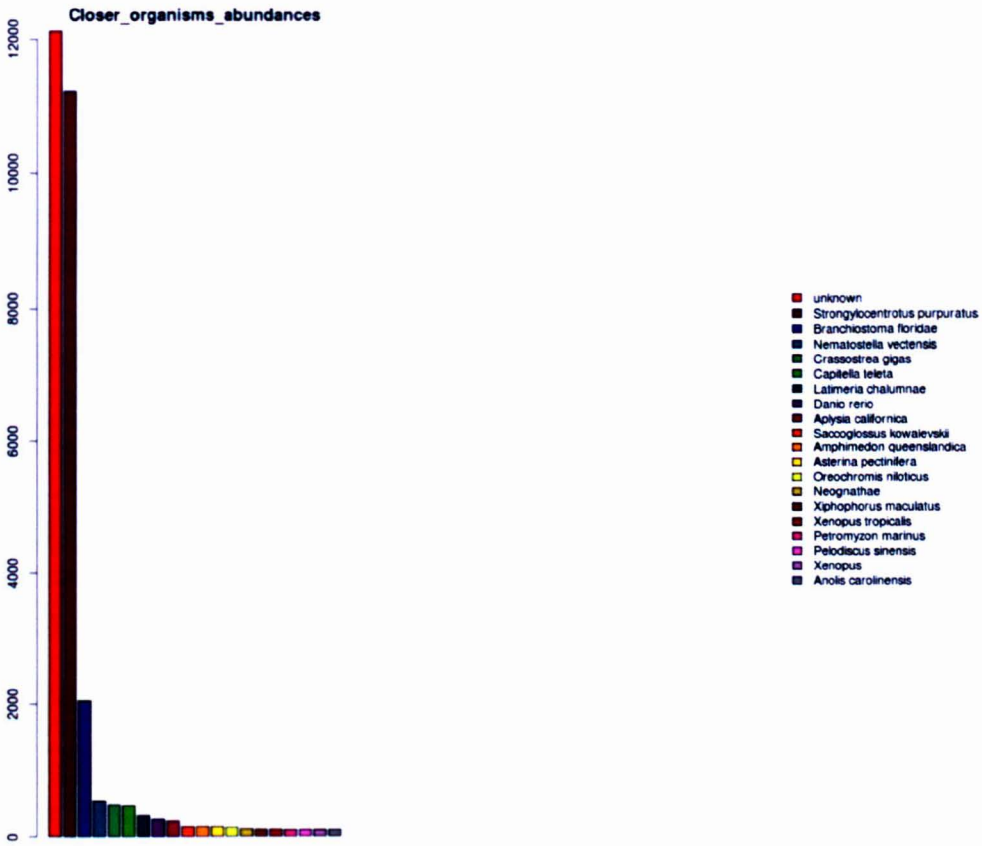
(obtained with probability major than: 0.95 and maximum length of the orf: 100)


Biological_Process_percentages
0.0
0.2
0.4
0.6
0.8
1.0
1.2
RNA-dependent DNA replication
proteolysis
signal transduction
transmembrane transport
intracellular signal transduction
regulation of transcription, DNA-dependent
transcription, DNA-dependent
DNA integration
small GTPase mediated signal transduction
oxidation-reduction process
ATP catabolic process
translation
microtubule-based movement
metabolic process
ubiquitin-dependent protein catabolic process
carbohydrate metabolic process
intracellular protein transport
protein ubiquitination
protein folding
multicellular organismal development


Molecular_Function_percentages
0
1
2
3
4
5
ATP binding
zinc ion binding
nucleic acid binding
DNA binding
RNA binding
calcium ion binding
RNA-directed DNA polymerase activity
oxidoreductase activity
metal ion binding
protein serine/threonine kinase activity
sequence-specific DNA binding transcription factor
sequence-specific DNA binding
nucleotide binding
catalytic activity
GTP binding
G-protein coupled receptor activity
hydrolase activity
phospholipid binding
protein kinase activity
electron carrier activity


Cellular_Component_percentages
0
1
2
3
4
5
integral to membrane
nucleus
cytoplasm
membrane
intracellular
plasma membrane
microtubule
mitochondrion
ribosome
extracellular region
endoplasmic reticulum
microtubule associated complex
Golgi apparatus
cytosol
nucleolus
endoplasmic reticulum membrane
chromosome
transcription factor complex
cytoskeleton
mitochondrial inner membrane


Conserved_Domains_percentages
0.0
0.5
1.0
1.5
pfam00069 Pkinase Protein kinase domain.
pfam01576 Myosin_tail_1 Myosin tail. The myosin mo
pfam10174 Cast RIM-binding protein of the cytomatr
pfam12126 DUF3584 Protein of unknown function (DUF
pfam13868 Trichoplein Tumour suppressor, Mitostati
pfam07690 MFS_1 Major Facilitator Superfamily.
pfam02463 SMC_N RecF/RecN/SMC N terminal domain. T
pfam05557 MAD Mitotic checkpoint protein. This fam
pfam12796 Ank_2 Ankyrin repeats (3 copies).
pfam07679 I-set Immunoglobulin I-set domain.
pfam07714 Pkinase_Tyr Protein tyrosine kinase.
pfam00076 RRM_1 RNA recognition motif. (a.k.a. RRM
pfam00041 fn3 Fibronectin type III domain.
pfam13923 zf-C3HC4_2 Zinc finger, C3HC4 type (RING
pfam05483 SCP-1 Synaptonemal complex protein 1 (SC
pfam00271 Helicase_C Helicase conserved C-terminal
pfam13639 zf-RING_2 Ring finger domain.
pfam00400 WD40 WD domain, G-beta repeat.
pfam00078 RVT_1 Reverse transcriptase (RNA-depende
pfam05701 DUF827 Plant protein of unknown function



Annocript 2.0 - Copyright of Bioinformatics Lab SZN Naples

Thu Nov 7 09:57:04 2013

Annexure 2

Candidate lincRNAs in zebrafish islet cells

Name	Locus	Islet	Embryo	Closest	Orientation	Distance
TCONS_00020484	chr2:39641763-39659299	48	0	*epha4a*	3prox	112719
TCONS_00026726	chr23:23277206-23280244	91	13	*samd11*	5prox	34608
TCONS_00025575	chr23:9414468-9415433	142	21	*acss2*	3prox	19134
TCONS_00018959	chr2:29984139-29984988	442	89	*tmem70*	3prox	16961
TCONS_00023208	chr21:28340462-28346900	96	22	*ppp3ca*	div	7661
TCONS_00030553	chr3:2306426-2309852	108	37	*si:dkeyp-52c3.2*	con	6800
TCONS_00023841	chr21:28340462-28346900	123	25	*ppp3ca*	div	6006
TCONS_00010097	chr14:18736495-18738047	256	18	*slbp*	5prox	3623
TCONS_00010098	chr14:18736495-18738047	260	18	*slbp*	5prox	3623
TCONS_00010099	chr14:18736495-18738047	256	18	*slbp*	5prox	3623
TCONS_00019413	chr2:9535344-9536956	132	48	*dvl3a*	5prox	2367
TCONS_00021096	chr20:52885313-52886283	249	39	*eef1db*	5prox	1441
TCONS_00002359	chr1:32881285-32881826	57	17	*prkx*	5prox	1338
TCONS_00028115	chr24:10218094-10310012	344	25	*myca*	3prox	1317
TCONS_00032543	chr4:7443862-7447881	478	152	*ERC1 (2 of 3)*	con	975
TCONS_00032542	chr4:7443862-7447881	484	152	*ERC1 (2 of 3)*	con	962
TCONS_00033332	chr4:7443862-7447881	488	152	*ERC1 (2 of 3)*	con	962
TCONS_00033333	chr4:7443862-7447881	489	153	*ERC1 (2 of 3)*	con	962
TCONS_00040802	chr8:17129287-17132842	2255	522	*tor3a*	5prox	885
TCONS_00040803	chr8:17129287-17132842	2238	517	*tor3a*	5prox	884
TCONS_00033298	chr4:2214772-2215137	77	13	*nap1l1*	3prox	876
TCONS_00021352	chr20:23068881-23070783	684	134	*usp46*	div	531
TCONS_00043074	chr9:15843641-15846550	761	300	*fn1*	div	145445
TCONS_00039553	chr7:62448205-62448548	130	29	*acox3*	div	10870
TCONS_00023842	chr21:28340463-28346900	34	4	*ppp3ca*	div	5949
TCONS_00042198	chr9:35342907-35344637	205	54	*cd247*	div	2728

Name: Assigned transcript name for the lincRNAs

Locus: Genomic location

Islet: Number of raw sequencing reads mapped on the transcript in islet cells

Embryo: Number of raw sequencing reads mapped on the transcript in Embryo cells

Closest: Closest coding gene in the zebrafish genome-wide

Orientation: Orientation of the lincRNA with respect to the closest coding gene

5Prox: 5' proximal; 3prox: 3' proximal; div: Divergent; cov: Convergent

Distance: Distance from the closest coding gene